# LONGITUDINAL ASSESSMENT OF DAILY ROUTINE UNIFORMITY IN A SMART HOME ENVIRONMENT USING HIERARCHICAL CLUSTERING

A Thesis

by

PRAKHAR MOHAN

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

| | |
|---|---|
| Chair of Committee, | Theodora Chaspari |
| Co-Chair of Committee, | Ryan Changbum Ahn |
| Committee Member, | James Caverlee |
| Head of Department, | Scott Schaefer |

May 2020

Major Subject: Computer Science

ABSTRACT

The gradual decline in routine patterns is a major symptom of early-stage dementia, therefore an unobtrusive real-life assessment of the elder's routine can potentially be of significant clinical importance. This research focuses on the assessment of changes in a person's daily routine using longitudinal data recorded from a network of non-intrusive motion sensors in a smart home environment. We propose to identify repeating patterns in a person's daily routine over the span of multiple days using hierarchical clustering algorithms, which allow us to disregard noisy signal patterns and various confounding factors that contribute to the momentary variability of the sensor data. We have evaluated our proposed algorithm on both synthetic and real-world data recorded in the span of 50-100 days from four elderly adults. Our results indicate that the proposed hierarchical clustering approach can more reliably quantify the degree of routinness compared to baseline approaches that compare the routines of two consecutive days or capture variations in the occurrence of recognised activities.

## ACKNOWLEDGMENTS

I would like to begin by expressing my heartfelt gratitude to my advisors Dr. Theodora Chaspari and Dr. Ryan Changbum Ahn who helped me tremendously throughout this research journey. Dr. Chaspari and Dr. Ahn have been the most patient, helpful and inspiring mentors who guided me at every step towards the successful completion towards this thesis. Their constant support and suggestions in this research played a key role in improving this work. I would also like to thank Dr. James Caverlee for serving on my committee and his constant support.

I would like to thank my fellow researchers at HUBBS and SSC research groups: Bogyeong Lee, Hoonyong Lee, Jinwoo Kim, Vansh Narula, Projna Promita, and Ehsanul Haque Nirjhar for their constant support and encouragement. I also want to thank my friends Nikhil Bhat and Manish Patel for being my emotional support pillars throughout this journey.

Finally, I am thankful for my parents who have always taught me the value of education and hard work. It is due to their sacrifices, undying love, and believe that I have succeeded in achieving this goal.

## CONTRIBUTORS AND FUNDING SOURCES

The thesis committee for this work include Dr. Theodora Chaspari (Chair) from the Department of Computer Science, Dr. Ryan Changbum Ahn (Co-Chair) of the Department of Construction Science and Dr. james Caverlee of the Department of Computer Science at Texas A& M University.

The datasets for this research were generated with the help of Bogyeong Lee who also helped in pre-processing the real-life data used for this research. All other work conducted as part of this thesis was completed by the author independently.

This research was supported by a grant (19CTAP-C128499-03) from the Technology Advancement Research Program funded by the Ministry of Land, Infrastructure, and Transport of the Korean government.

## NOMENCLATURE

| | |
|---|---|
| WHO | World Health Organization |
| MCI | Mild Cognitive Impairment |
| ADL | Activities of Daily Living |
| PP | Poincare Plot |
| SEEDS | Super-pixels Extracted via Energy-Driven Sampling |
| RF | Radio Frequency |
| PIR | Passive Infrared |
| RMSE | Root Mean Square Error |
| MAE | Mean Absolute Error |

TABLE OF CONTENTS

LIST OF FIGURES

FIGURE Page

LIST OF TABLES

# 1. INTRODUCTION AND LITERATURE REVIEW

## 1.1 Mild Cognitive Impairment (MCI)

In the last few decades, there have been a significant increase in cognitive diseases such as dementia and depression among our elderly population. The World Health Organization (WHO) reported a statistics of around 50 million cases of dementia worldwide during 2015 with 63% patients living in low or middle income countries where medical facilities are not readily available [3]. Alzheimer's Association reported an estimate of 5.8 million cases in the United States as of 2019 of which 81% patients are of age 75 or older [4]. These cognitive diseases are of slow progressive nature and have adverse impact on mental and physical functioning causing forgetfulness, memory loss, reduced concentration and reduced ability to perform basic day to day activities, called activities of daily living (ADL), such as bathing, eating, and cooking [5, 6]. Mild cognitive impairment (MCI) is the stage between the expected cognitive decline of normal aging and the more serious decline of dementia. A meta analysis found that 38% patients with mild cognitive impairment (MCI) developed dementia over a period of five years [4]. Although current advancements in medical research has made it possible to clinically diagnose MCI [7, 8], the slow progression their symptoms render such diseases often difficult to identify during early stages leading to a total ignorance of the disease development.

## 1.2 ADL Assessment and Smart Home Environment

The relation between ADL patterns and cognitive diseases is a well studied topic [9, 10] and results indicate that an assessment of ADL patterns and daily routines is a crucial step towards the diagnosis of these diseases [11, 12, 13]. Traditionally, activity assessment for clinical purposes was performed by closely observing a person doing specific activities in a controlled environment or through surveys reported by self or a caretaker[14, 15]. The clinical observations are performed once every few month in order to track the health status or progression of diseases. These methods usually take a long time, the clinical settings are often different than the real settings and the subjects tends to be more careful while performing regular tasks. The self reported observations are generally biased and erroneous due to ignorance of symptom details because of the lack of knowledge about the disease [16]. These problems are more prevalent among older adults living alone who do not have access to proper care facilities.

The analysis of ADL patterns during the early stages of these diseases require long-term monitoring of the occupants' behaviour which is often difficult to perform in a clinical environment. Modern advancements in smart home technologies can leverage sensor networks and smart devices, such as wearable sensors, motion capture devices, audio and video sensors [17, 18, 19, 20], to collect rich continuous data and enable the longitudinal health and behavior monitoring. The long term assessment of declining routine patterns in a smart home environment can therefore provide insightful information on the development and progression of chronic cognitive diseases [21].

Various sensor components used in smart homes can be broadly categorized into intrusive and non-intrusive sensors. Intrusive sensors are highly sophisticated sensors, such as microphones, cameras, and wearable sensors which provide a diverse set of rich audiovisual and physiological data. These sensors can be extremely useful to model and contextualize ADL, but often raise privacy concerns and cause individual discomfort preventing their wide adoption. As a potential solution to these issues, non-intrusive sensors, such as a networks of motion and touch sensors installed on various home objects, are widely used in smart home environments [22, 23]. While the information obtained from these sensors is limited and noisy, if combined with emerging signal processing and data analytics algorithms, it has the potential to yield valuable insights towards the assessment of human daily routines.

### 1.2.1 Challenges in daily routine assessment

Smart homes with a diverse set of sensors provides a great platform for long-term assessment of daily routines. Yet, this endeavor comes with many challenges.

#### 1.2.1.1 *Challenge 1: Sensor noise*

Data collected from the smart home sensors are often affected by multiple confounding factors, such as low-frequency noise from electronic interference, signal energy spill over among adjacent sensors, and inaccurate sensor measurements in proximal room locations. The noise in these sensors results in inaccurate readings at the sensor level and might have huge impact on the assessment experiment. Another situation often encountered in activity data is the inaccurate sensor measurements in proximal room locations. This situation arises due to an activity at the intersection of multiple sensors which often introduce huge amount of noise in the data of certain activities. A similar situation is shown in Fig. 1.1 where multiple sensors are actuated as the person moves from point 'a' to 'b'. To reduce the effects of sensor noise, we use median filtering on the raw data.

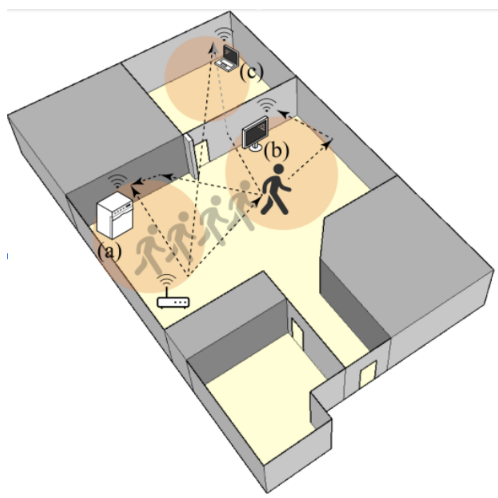



Figure 1.1: Multiple sensors actuating while person moves through an intersection.

#### 1.2.1.2 *Challenge 2: High dimensional data*

Smart homes provide massive amount of data generated through multiple sensors. These sensors capture data at high frequencies and the data generated through different sensors is often asynchronous. As a person performs a diverse set of activities every day, the activity data obtained through these sensors is often unlabeled and is difficult to annotate. Data annotation is generally done by asking the patients or caretakers to record every activity or by attentively labeling the high dimensional data manually. This is a strenuous task and often results in human errors. Also, asking to record every activity leads to unwanted displeasure among the patients or caretaker. This thesis will leverage an unsupervised algorithm to assess the daily routine of a person. Also, for a detailed evaluation of our approach, we propose a synthetic data with labeled activities generated by systematically controlling variations in a predefined daily routine.

#### *1.2.1.3 Challenge 3: Inconsistencies in daily routines*

Another common problem in routine assessment is to identify the activities that are being performed on regular basis and contribute to shaping one's routine patterns. Out of the various activities one performs in day-to-day living, only a few of them form part of the routine. The daily routine often varies across days; for example, the routine of different weekdays may be different, or the morning routine over weekdays might be similar, but evening routine might be different. These inconsistencies in routine patterns further add to the complexity of the considered problem. Fig. 1.2 demonstrates such a scenario where the routine for first half on day 1 and 2 is similar while the second half of the two days have different routines and the second half of day 1 and 3 is similar. The figure also demonstrate some activities which do not form a part of daily routines. In this report, we refer small non-routine contributing activities as noise activities within the daily routines. To capture inconsistencies in daily routines this thesis presents a hierarchical clustering algorithm to identify similarly repeating activity patterns over a span of multiple days.

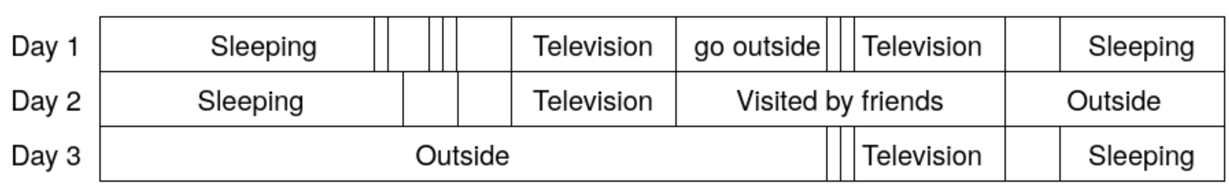



Figure 1.2: A hypothetical example of inconsistent activity routine

## 1.3 Prior Work

Early identification or progression of cognitive diseases in smart homes is often done by assessing a set of recognised activities [1, 24] or by directly processing the sensor information [23]. Activity recognition in smart home environment though intrusive as well as non-intrusive sensors have been a well studied topic. Activity recognition through intrusive sensors [25, 26, 27, 28], have demonstrated promising results. Even though these sensors provide rich information, they often raise the concerns related to privacy invasion and personal discomfort and are not readily acceptable for long term monitoring. Activity recognition through non-intrusive sensors is a hot research topic [29, 22, 30]. Non-intrusive sensors are widely used in smart homes and provide a diverse set of information such as appliance usage, furniture usage, presence and motion information.

Urwyler et. el. in [1] uses Poincare plots to assess the uniformity of daily routines through a set of recognized activities. This work uses a rule base ad-hoc classifier proposed in [29] to detect and classify a set of ADLs. The classifier is based on the assumption that some activities are performed at a specific time, for a specific duration and has a specific feature distribution. Fig. 1.3 shows the detected activities for two days. Similar activities detected on two consecutive days are then paired and plotted on a Poincare plot (PP) with respect to the start time of the activity.

The routines are assessed by fitting a best fit ellipse on the Poincare plot and evaluating the ellipse centroid and the two axis. This is shown in Fig. 1.4. Even though Poincare plots give a

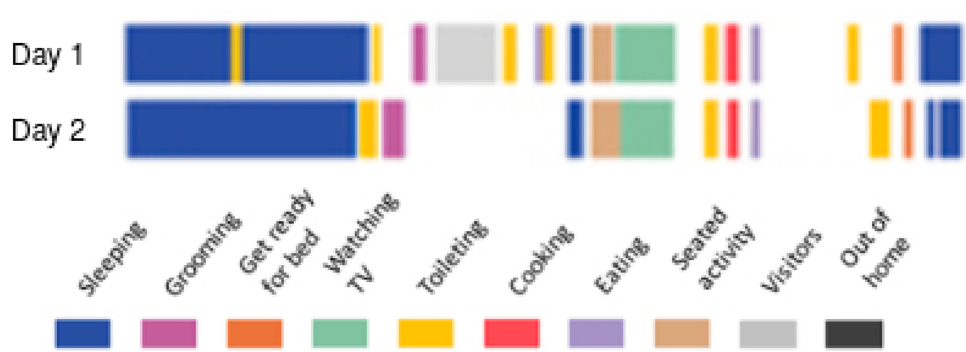
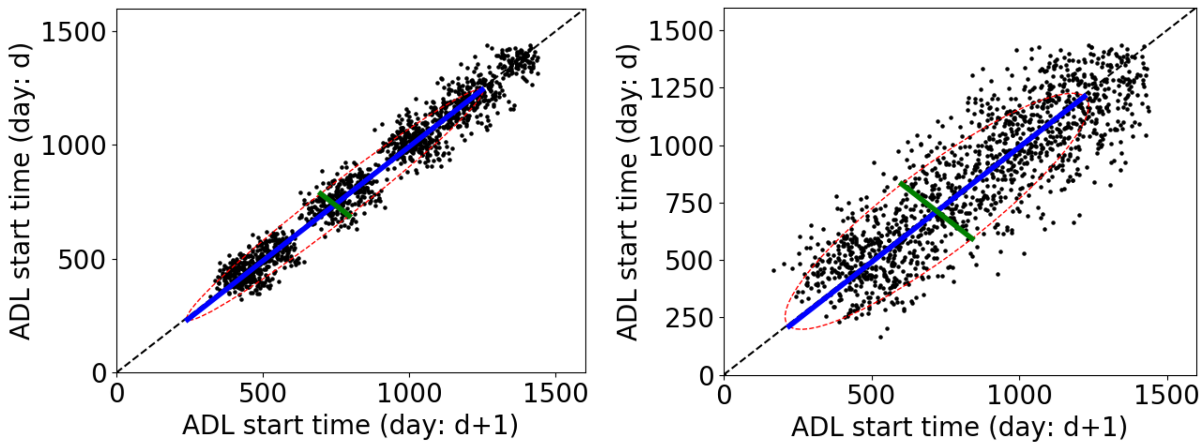



Figure 1.3: Recognised set of activities for two consecutive days. Adapted from Urwyler et. at. [1] with permission from Creative Commons Attribution 4.0 International License.




(a) More uniform routine



(b) Less uniform routine

Figure 1.4: Poincare plots

good comparison metric, this approach suffers many challenges. The daily routines are inconsistent and activities may not be repeated in a similar fashion every day. This makes pairing of similar activities extremely difficult. Also, the method lacks any standard way of pairing similar activities and require human efforts making the task more difficult.

Alberdi et. el. in [24] propose the use of Gestalt Sequence Matching [31] to compute the similarity between the routine of two days. The method uses 'AR' activity recognition algorithm [32] to detect ADLs. The recognised activities for a day are then represented as a routine sequence and compared with the sequence of the next day. Gestalt sequence matching uses the concept of longest common sub-sequences (LCS) to compute the similarity between two sequences. The similarity is calculated using the equation given below

$$Similarity = \frac{2 * \sum_i^n |LCS_i|}{|Seq_1| + |Seq_2|} \tag{1.1}$$

where the two sequence have $n$ longest common sub-sequences and $|\cdot|$ is the length of the sequence. The similarity computation can be understood through an example presented in fig. 1.5 where there are two common sub-sequences ('WIKIM', 'IA') found in the two sequences. The similarity is calculated using equation 1.1 as $2 * (5 + 2)/(9 + 9) = 0.78$. This method heavily

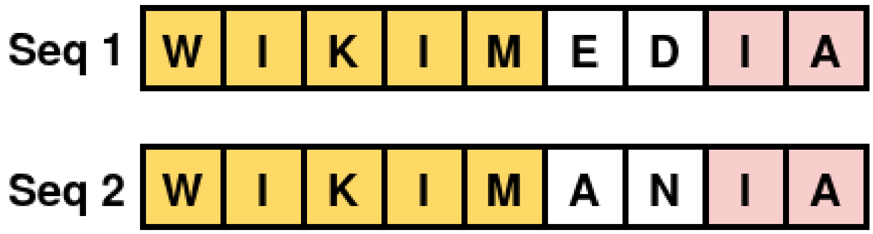


Figure 1.5: Example for Gestalt Sequence Matching [2].

suffers from the variations in similar activities within the daily routines. Presence of noise activities splits a longest common sub-sequence into smaller sub-sequences which reduces the similarity between the routine sequences of two days. Also, the presence of long duration activities such as sleeping, introduces bias and reduces the impact of smaller activities in the computation of similarity between daily routines.

A drawback of routine assessment through recognised activities is that activity recognition models vary according to different settings and therefore requires huge amount of training data for every other subject. Another challenge in these approaches is the lack of annotated data. Annotation of high dimensional data obtained from non-intrusive sensors is a difficult task and is not always possible. To avoid these challenges, researches propose to assess the routine by directly processing the sensor data rather than first recognising key activities. These approaches use various methods like sensor fusion, auto-encoders and Markov Chain models to assess the daily routines.

Sharma and Ghose in [23] proposes the use of autoencoders to represent daily routines as fixed length encodings. They use multiple sensors which capture information at high frequencies thus generating high dimensional data every day. The sensor data for each day is fed to an auto-encoder to reduce the dimensions and generate fixed length encoding for the routines of each day. The encodings of consecutive days are then compared to evaluate the changes in the daily routines. This method is more robust to presence of noise activities and variations in the occurrance of the activities but are still highly effected by the inconsistencies in the daily routines.

Markov models is another approach which researchers explored to represent daily routines. Shirin et. el. in [33] uses Marcov chain to model the routines in order to detect the unusual activity patterns of people with dementia. Their approach is targeted towards finding aberrations or sudden changes in the daily activity patterns of the people. These methods provide valuable information on short term changes or abnormalities in activity routines and may not give insightful results for the long-term assessment of daily routines in early stage cognitive impairments.

## 1.4 Summary of Proposed Work and Contributions

The aim of this research is to identify gradually declining routine uniformity which might indicate the development of progressive cognitive diseases using non-intrusive in-home monitoring. We exploit the idea that progression of cognitive diseases is characterised by a gradual decline in the uniformity of activity patterns. We propose a novel approach to identify repeating patterns of similar ADLs across several days of one's life using hierarchical clustering and region growth algorithms. The proposed algorithm divides the activity data of each day into smaller homogeneous segments and captures the uniformity of routines by assessing the repetition of these segments over multiple days. We focus on unlabeled activity data obtained though a network of non-intrusive wireless radio-frequency (RF) motion sensors installed at various locations within elderly individual's home and is evaluated through the acquired real-life data, as well as carefully generated synthetic data for the purposes of our problem. Our approach is evaluated against previously proposed routine assessment approaches that rely on day-to-day similarity metrics (e.g., sequence matching and assessment using Poincare Plots). Results from our analysis indicate that the proposed approach leveraging hierarchical clustering and region growth algorithms is more robust to inconsistencies in the daily routines and can capture variations in the uniformity of routines in synthetic and real data. We show that our approach is able to capture routine patterns that repeat frequently and infrequently and also capture the variations in the repetition of these patterns.

## 1.5 Research Objectives

### 1.5.1 Research Aims

This thesis attempts to answer the following main research questions:

1. **Is there a way to find activity patterns repeating on multiple days?**

We propose to identify patterns within the daily routine that repeat over multiple days. As only a few activities contribute to the routine of a person, these repeating patterns are expected to represent the routine activities. To answer this question we split the problem statement into two sub-questions.

(a) **Can we divide the time series data for each day into smaller meaningful segments?**

We divide the time-series data for each day captured by motion sensors into smaller contiguous homogeneous segments. This is achieved by making necessary modifications in a widely used image segmentation algorithm called Super-pixels Extracted via Energy-Driven Sampling (SEEDS) [34].

(b) **How can we find similar segments repeating in a similar fashion?**

The segments achieved through SEEDS algorithms for multiple days are clustered together using hierarchical clustering algorithm. In doing so, similarly repeating patters are grouped together. These groups of similar segments are expected to represent the routine of the person.

2. **How to capture uniformity or variations in the activity patterns?** The uniformity of daily routines is assessed by evaluating the groups or clusters of similar segments though different measures.in the daily routines.

# 2. DATA DESCRIPTION

We evaluate our approach on two sets of data, a set of real-life data collected from elderly adults in their home environment, as well as synthetic data generated for the purposes of this problem with pre-defined levels of noise and randomness. Although synthetic data cannot replace real measurements, they provide us with a systematic and principled way to evaluate our approach and partially address the absence of labels in the real dataset.

## 2.1 Real-Life Data

We collected data from four elderly adults living in single occupant houses. Table 2.1 provides some details about the four subjects as well as the number of days for which the data is collected. Subject 1 is of 72 year of age with a minor depression. Subject 2 and 3 are healthy participants with age of 73 and 75 respectively. Subject 4 is of 82 years of age and had been diagnosed with Parkinson's disease. Subject 1 and 3 perform frequent long visits outside the house thus they have lower number of room transitions per day compared to subject 2 and 4.

| Sub | Age | Mental Health | Details | Days |
|---|---|---|---|---|
| 1 | 72 | - Self reported depression<br>- Minor medication | - Frequent outside visits in noon<br>- Location transitions (avg. 52) | 45 |
| 2 | 73 | - Healthy subjects | - Evening walks, often at home<br>- Location transitions (avg. 97) | 116 |
| 3 | 75 | | - Overnight stays in other house<br>- Location transitions (avg. 50) | 64 |
| 4 | 82 | - Parkinson's Disease | - Stays home, weekly visits to hospital<br>- Location transitions (avg. 74) | 62 |

Table 2.1: Subject description for collected data.

The data is collected through a network of non-intrusive wireless motion sensors installed at various locations within each individual's house [35]. Fig. 2.1 shows the sensor installation within the house of Subject 1. These sensors provide information like timestamp and motion coordinates

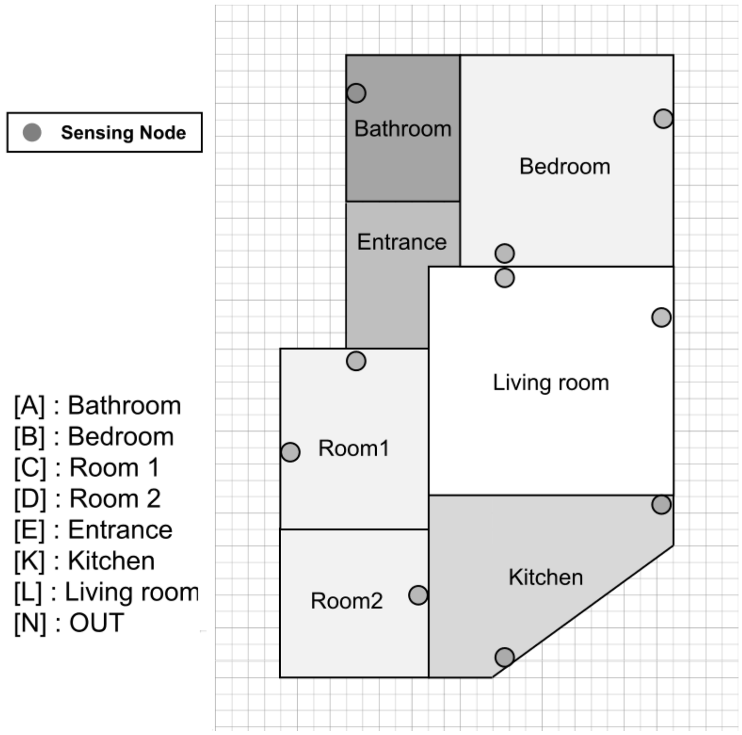



Figure 2.1: RF motion sensors at various locations within the house of Subject 1.

with respect to the center of the house. These coordinates are then mapped to the corresponding room/location within the house. Throughout the period of data collection, the subjects were allowed to follow their usual routine without any restriction or obtrusion. The raw data was pre-processed through median filtering to reduce high noise present in the sensor data. The final data that we are working on include a time-series of room identities (IDs) sampled every minute. Fig. 2.1 shows the IDs corresponding to the different rooms within the house of the subject.

### 2.2 Synthetic Data

As data collection is a resource-intensive process and it is difficult to obtain fully labelled data in real-life, we further attempt to evaluate our results on a set of synthetic data generated for the purposes of this problem. We generated synthetic data by pre-defining a sequence of activities and systematically introducing multiple levels of noise and randomness. Table 2.2 provides the predefined routine used to generate the synthetic data. In order to make the synthetic data as consistent as possible to the real-life data, we further pre-define room IDs for each activity. Table

| Start Time(h:m) | Duration (min) | Activity | Noise Activities (n% of N instances) |
|---|---|---|---|
| 00:00 | 420 | Sleeping | 3% |
| 07:00 | 30 | Personal Hygiene | 6.5% |
| 07:30 | 90 | Cook/Eat | 15% |
| 09:00 | 180 | Entertainment | 15% |
| 12:00 | 90 | Cook/Eat | 15% |
| 13:30 | 180 | Sleeping | 3% |
| 16:30 | 30 | Personal Hygiene | 6.5% |
| 17:00 | 120 | Outside | 3% |
| 19:00 | 90 | Cook/Eat | 15% |
| 20:30 | 150 | Entertainment | 15% |
| 23:00 | 60 | Sleeping | 3% |

Table 2.2: Predefined synthetic routine for a day.

2.3 provides the the room IDs corresponding to each activity of the predefined routine. Each segment of sleeping and personal hygiene activity can occur in any one of their respective locations (R1,R2,R3 and T1,T2), selected randomly each time, while others can occur in a combination of one or more locations as provided in the corresponding rows.

The randomness in synthetic routines is introduced using three variables. The first variable is the standard deviation ($\sigma$) in the duration of activities as given in Table 2.2. The $\sigma$ is the $X$ percentage of activity duration where x $\in$ {5, 10, 15, 20, 25, 30}. A higher value of $X$ results in lesser uniform routines. The second variable is the number of total noise activities ($N$) allowed within a day where $N \in \{20, 30, 40\}$. The last column of the Table 2.2 defines the number of noise activities allowed in each corresponding activity as the $n^{th}$ percentage of $N$. The third variable P controls the consistency or order of activities. With a probability of $P$, the activities are performed in order of Table 2.2 and with a probability $1 - P$ the activities are not performed in order. To have routines with different consistencies, $P$ is varied as $P \in \{0.9, 0.7, 0.5, 0.3\}$. We present three types of ADL routines based on the degree of noise and randomness introduced.

| Activity | Room/Location |
|---|---|
| Sleeping | Room (R1)<br>Room (R2)<br>Room (R3) |
| Personal Hygiene | Washroom (T1)<br>Washroom (T2) |
| Cook/Eat | Living room (L), Kitchen (K) |
| Entertainment | Living room (L) |
| Outside | Outside (O) |
| Noise Activities | R1, R2, R3, T1, T2, L, K, O, E |

Table 2.3: Activities and corresponding rooms and locations.

### 2.2.1 Synthetic Routine Type 1

This routine type consists of activities occurring in the same order as provided in Table 2.2. The duration of each activity is generated randomly using a Gaussian Distribution with mean $D(a)$ and deviation $\sigma$, where $D(a)$ is the actual duration of the activity $a$ in Table 2.2 and $\sigma$ is the $x$ percentage of D(a) where x ∈ {5, 10, 15, 20, 25, 30}. The uniformity of the routines decreases with increasing value of x. Each activity is then assigned a combination of locations according to Table 2.3 and as explained above.

### 2.2.2 Synthetic Routine Type 2

Synthetic routines in this class are generated by randomly introducing noise activities in Type 1 routines. The number of noise instances $N$ each day varies within the set $N \in \{20, 30, 40\}$ where $N$ is constant for all days in a particular routine. For each activity within a day, the number of noise activities is then selected randomly between 1 to $n\%$ of $N$, where $n$ for each activity is provided in Table 2.2. The uniformity for Type 2 routines decreases with increasing standard deviation in the duration of activities as well as increasing instances of noise activities.

### 2.2.3 Synthetic Routine Type 3

Unlike Type 1 routines, Type 3 routines depict inconsistencies in the order of occurrence of activities i.e. the activities may or may not occur in order. The activity at a given time is decided such that with a probability $P$ the selected activity will be in order and with $1 - P$ probability it will not be in order. Each activity is then assigned a combination of locations as explained above. With lower P, the activities are more probable to be out of order thereby reducing the consistency of the ADL routine. Type 1 patterns are a special case of Type 3 patterns where P=1 where the probability of not following the order is 0.

# 3. METHODOLOGY *

Our proposed approach of assessing ADL routines is based on two assumptions. The first is that out of various activities performed on day-to-day basis, only a few activities contribute to a person's daily routine and second is that most of the activities are associated with a specific distribution of locations within the house. For example, "Sleeping" activity is associated with the bedroom, while "Personal Hygiene" activity is associated with the washroom.

## 3.1 Image-based representation of daily activity patterns over time

We propose to arrange the sequence of activities for multiple days in a 2-dimensional matrix, where the y-axis (rows) and x-axis (columns) denote the day and time, respectively. Therefore, each pixel p(x,y) corresponds to the information or features of the activity being performed on day $y$ and time $x$. Our motivation behind representing the routine data as a 2-dimensional structure yielded by the fact that a completely uniform routine (i.e. all activities start and end at the same times) will follow the same vertical patterns across many days, therefore if we stack each day's activity sequence on top of each other, the segments of the same activities will completely overlap. Fig. 3.1 demonstrates the image representation of a completely routine where each color in the image depicts a unique activity.

On the contrary, the 2-dimensional representation of a non-uniform routine would be more varying and would not follow distinct vertical patterns in time. This can be observed from synthetic routines of Type 1 (Fig. 3.2), Type 2 (Fig. 3.3) and Type 3 (Fig. 3.4) for different values of variables X, N and P. Representing data as a 2-dimensional matrix provides us with the flexibility of using a large range of image processing tools to cluster together similar activities spanning a large temporal range.

*Parts of this chapter are modified from the CRC'20 paper [36] with permission of ASCE.

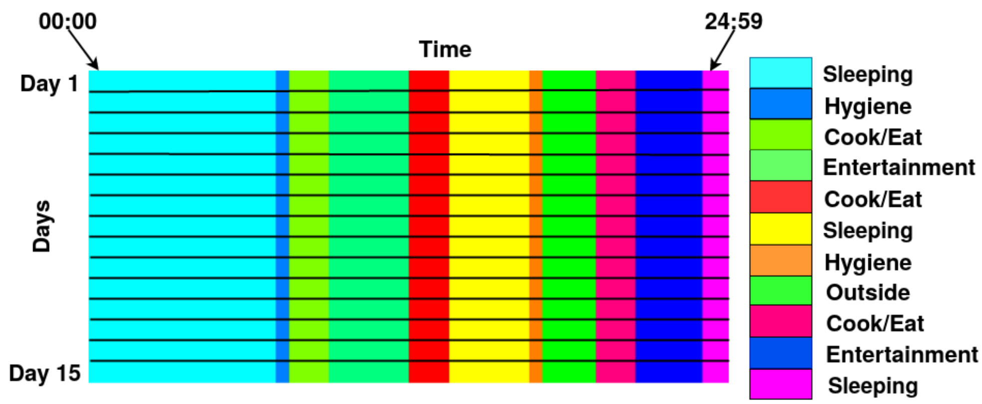
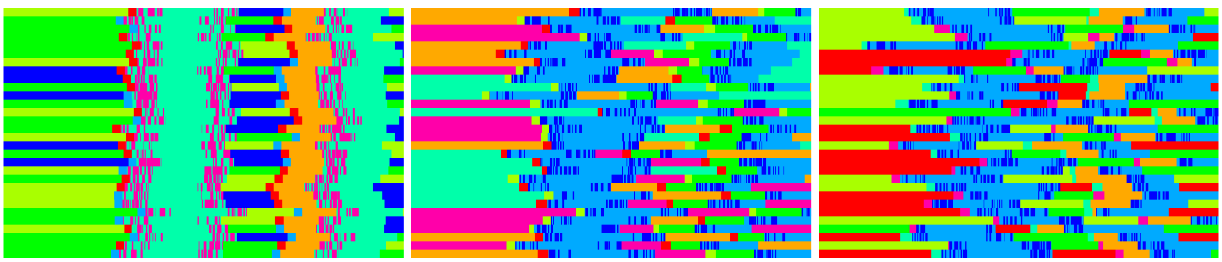
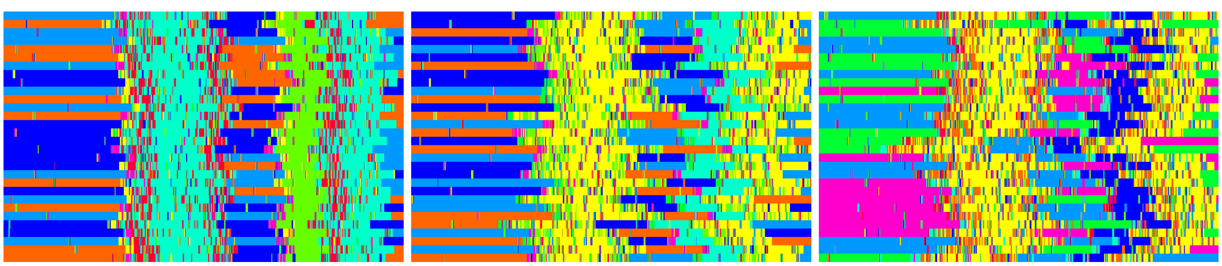



Figure 3.1: Completely uniform routine represented as an image.



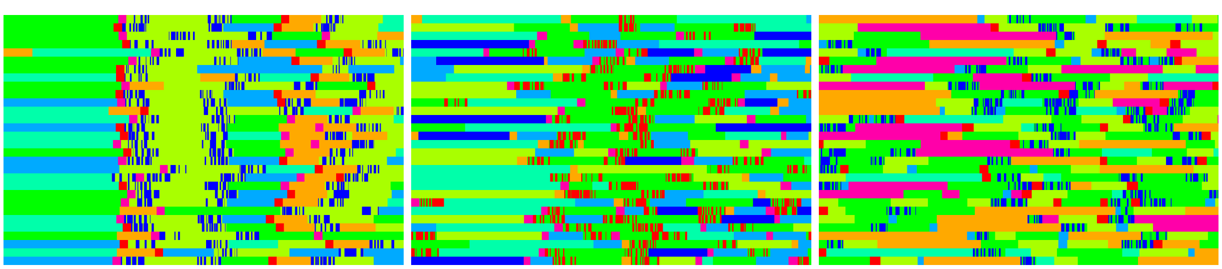
(a) X = 5 (b) X = 20 (c) X = 30

Figure 3.2: Representation of Type 1 synthetic routines.



(a) X = 5, N = 20 (b) X = 15, N = 30 (c) X = 15, N = 40

Figure 3.3: Representation of Type 2 synthetic routines.



(a) X = 5, P = 0.9 (b) X = 20, N = 0.7 (c) X = 20, N = 0.3

Figure 3.4: Representation of Type 3 synthetic routines.

## 3.2 Identifying activity segments from location-based data

This section answers our first research question on how can we divide the time series data for each day into smaller meaningful segments. An individual performs a diverse set of activities on day-to-day basis, including basic ADLs, instrumental ADLs, and more complex activities. Only a few of these activities actually contribute to the routine of the person. The high temporal granularity of sensor-based data introduces high levels of noise, therefore preserving a detailed time-sequence of location data might introduce unnecessary artifacts when computing similarity metrics that are used in clustering algorithms. For this reason, instead of performing clustering directly on the sensor-based time-series, we first group contiguous location-based measures with similar distribution into segments. Each segment contains a homogeneous distribution of locations and can be considered as an approximation of an activity during the day. In the following, we will describe the algorithm that was designed for this task.

We group location-based data into activity segments through a modified version of the Superpixels Extracted via Energy-Driven Sampling (SEEDS) algorithm [34]. SEEDS is commonly used for segmenting images into homogeneous regions through a hill climbing approach that iteratively refines the boundaries of the segmented regions [37, 38]. SEEDS algorithm works by dividing an image into smaller segments or blocks with each block subdivided into smaller blocks and so on until pixel level is reached thus forming a hill like structure. The top level segments (most coarse blocks) are then adjusted by moving the blocks at the boundary of two neighboring segments at each level. Fig. 3.5 demonstrates the working of SEEDS algorithm on an image. The initialization step shows the division of image into smaller segments. From left to right, each image shows the adjustment of boundaries of top level segments by moving blocks at different resolutions.

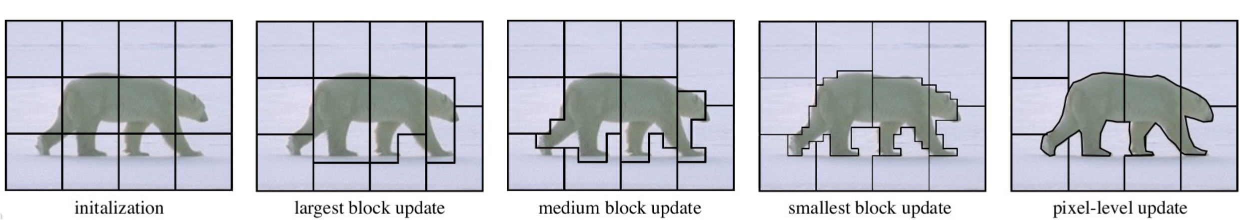



Figure 3.5: Demonstration of SEEDS algorithm on an image.

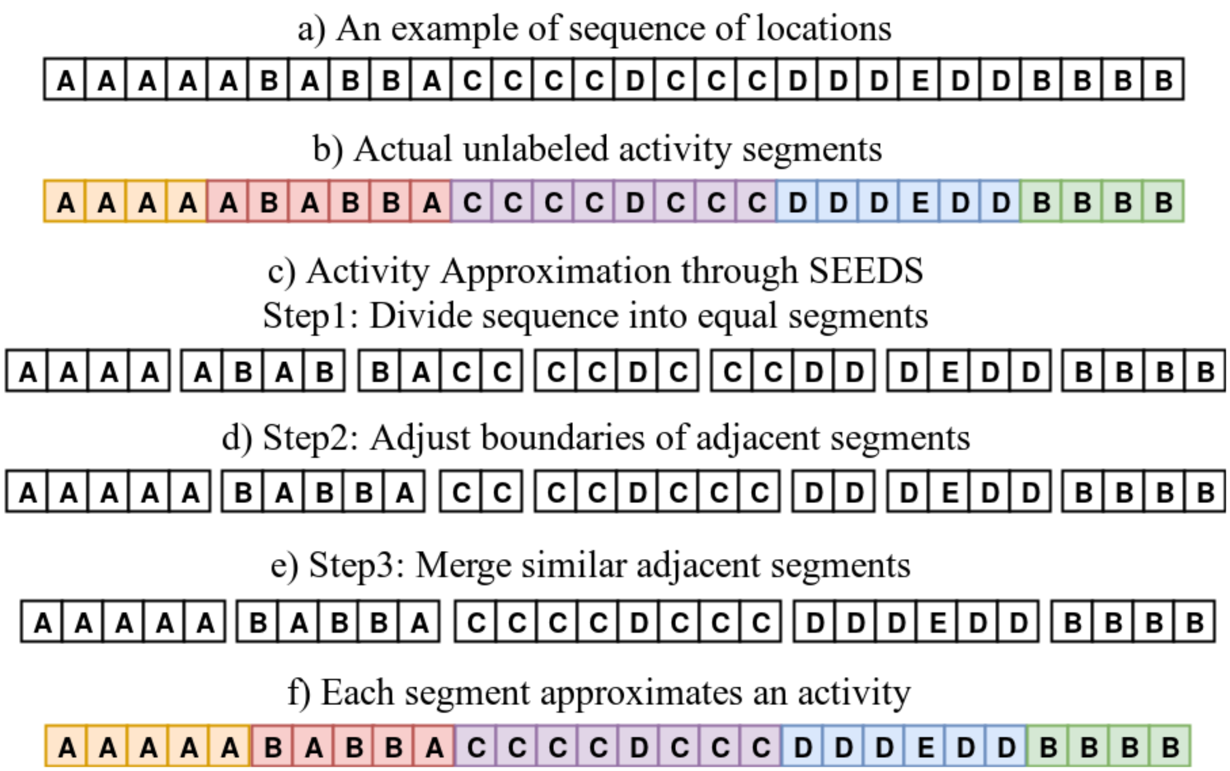



Figure 3.6: Example demonstrating the workings of the Superpixels Extracted via Energy-Driven Sampling (SEEDS) algorithm. (a) An example sequence of room identities; (b) Unlabeled activities and their location distribution; (c-f) Steps to approximate activities using SEEDS algorithm.

We modified the implementation of SEEDS to have a 1-dimensional sequence as an input, instead of the 2-dimensional image pixels. Fig. 3.6 provides an example on how a sequence of room IDs was grouped into homogeneous segments through the SEEDS algorithm. The input was the time-series depicting the location of the participant in the house, as opposed to the color intensity of the image proposed in the original image-based work [34]. SEEDS is initialized by dividing the location-based sequence into equal length segments. Each segment is further divided into equal length sub-segments, a procedure which is repeated until a segment consists only of one sample, thus forming a multilevel hill-like structure, as shown in Fig. 3.7.a. The segments at the bottom level consist of single locations. The upper level segments consist of groups of locations, which will be eventually refined to include contiguous location-based segments. Each segment created by the SEEDS algorithm is represented by a histogram of locations whose elements correspond to the unique locations present in the sensor-based time series.

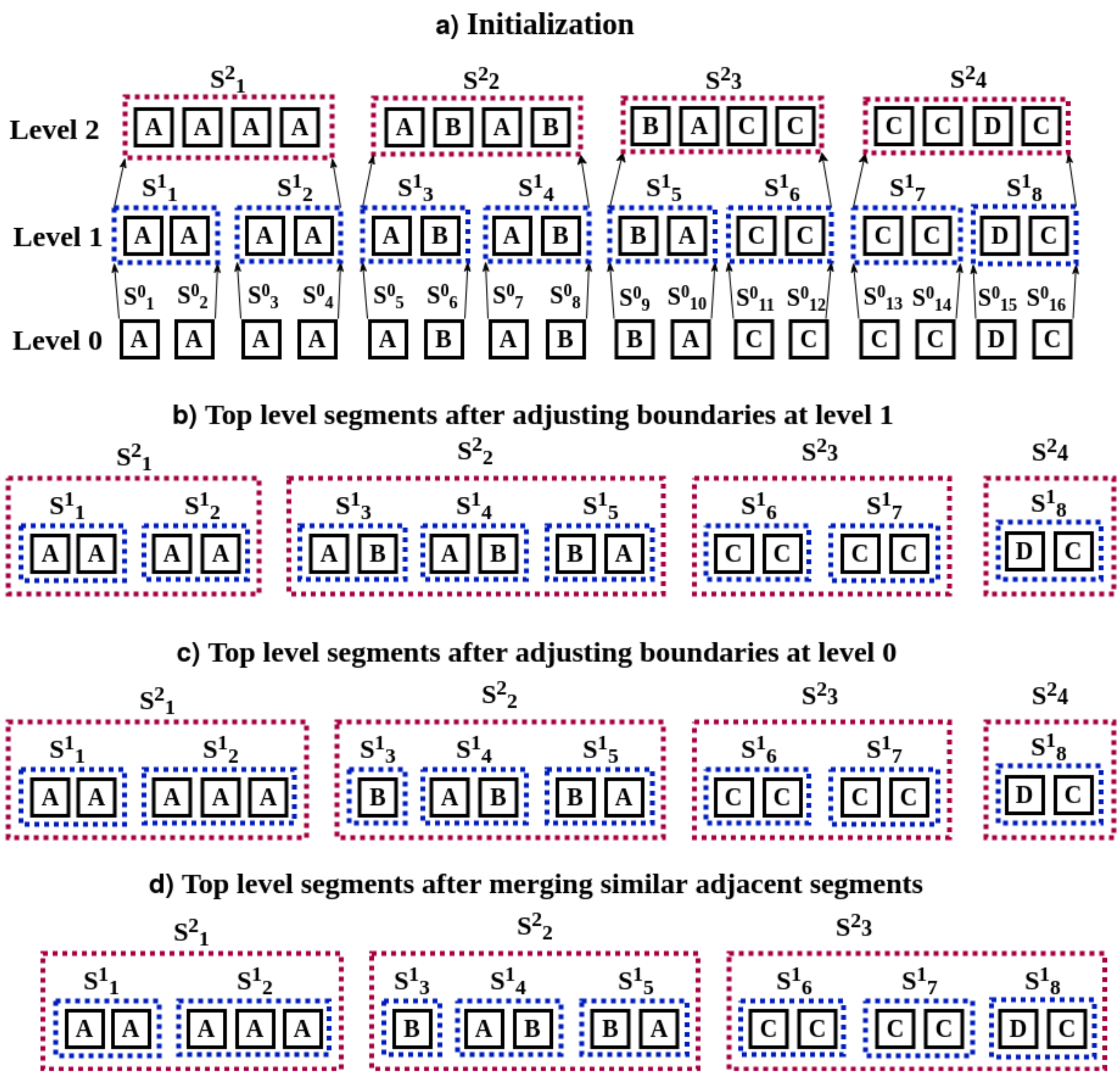



Figure 3.7: Different steps of SEEDS algorithm. a) Initialization into multilevel hill-like structure. b& c) Top level segments after adjusting boundaries at level 1 and 0 respectively. d) Top level segments after merging similar neighboring segments.

Let $\mathcal{L}$ be the set of unique locations. We define the location histogram $h^{loc}_{S^l_k}$ for the $k^{th}$ segment of level $l$, $S^l_k$, as follows:

$$h^{loc}_{S^l_k}[j] = \frac{1}{|S^j_k|} \sum_{i \in S^j_k} \mathbb{I}(i = j)\,, \quad for\ j \in \mathcal{L} \tag{3.1}$$

where $\mathbb{I}(\cdot)$ is the indicator function, such that $\mathbb{I}(A) = 1$, if $A$ is true, and $\mathbb{I}(A) = 0$, otherwise.

### 3.2.1 Adjustment of segment boundaries

Starting with the top most level and climbing down the hill, the boundary of two neighboring segments at the top level ($L$) is adjusted by moving the sub-segments at level $l$ which are at the boundary of the two segments. The boundary sub-segments is moved to the segment with which they have a higher Histogram Intersection value. The histogram intersection between two segments $S_1$ and $S_2$ is calculated as:

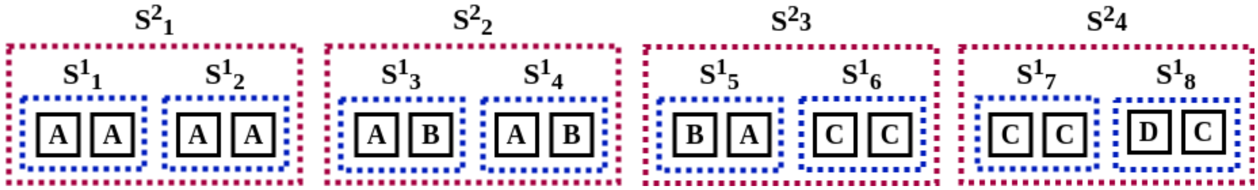
$$HI(h^{loc}_{S_1}, h^{loc}_{S_2}) = \sum_{j \in \mathcal{L}} min(h^{loc}_{S_1}[j], h^{loc}_{S_2}[j]) \tag{3.2}$$

If $S^L_a$ and $S^L_b$ are the two neighboring top level segments (e.g., $S^2_1$ and $S^2_2$, as shown in Fig. 3.7) and the boundary sub-segments of these two segments at level $l$ are $S^l_a$ and $S^l_b$ (e.g., $S^1_2$ and $S^1_3$, as shown in Fig. 3.7), then $S^l_a$ is moved from $S^L_a$ to $S^L_b$ if:

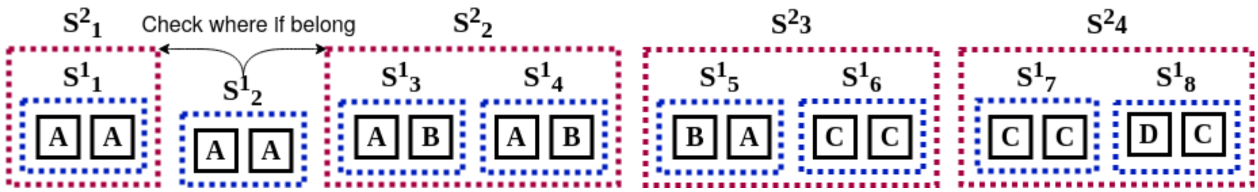
$$HI(h^{loc}_{S^L_b}, h^{loc}_{S^l_a}) - HI(h^{loc}_{S^L_a \setminus S^l_a}, h^{loc}_{S^l_a}) > \epsilon \tag{3.3}$$

where $S^L_a \setminus S^l_a$ is the segment $S^L_a$ without sub-segment $S^l_a$ and $\epsilon$ is a minimum threshold. After iteratively applying the adjustment of segment boundaries to all the levels, the top most level gives a resulting set of segments.

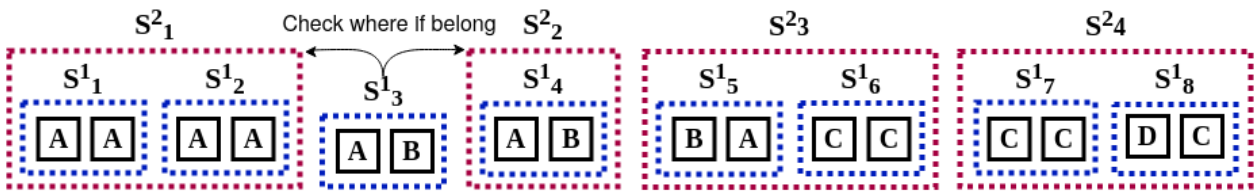
Fig. 3.7 b and c demonstrate the top level segments after adjusting the boundaries at level 1 and 0 respectively. After adjustment at level 1, it can be observed that $S^2_1$ remain unchanged, $S^1_5$ moved from $S^2_3$ to $S^2_2$, and $S^1_6$ moved from $S^2_4$ to $S^2_3$. Boundary adjustment at level 1 is followed by adjustment at level 0. It can be observed from 3.7.c that segment $S^0_5$ moved from $S^2_2$ to $S^2_1$ thus giving the resulting set of segments.



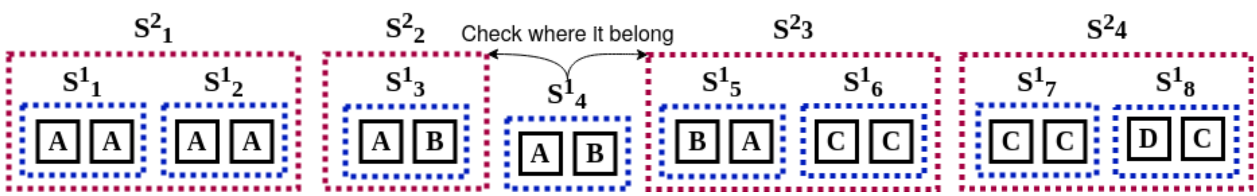

(a) Initialization



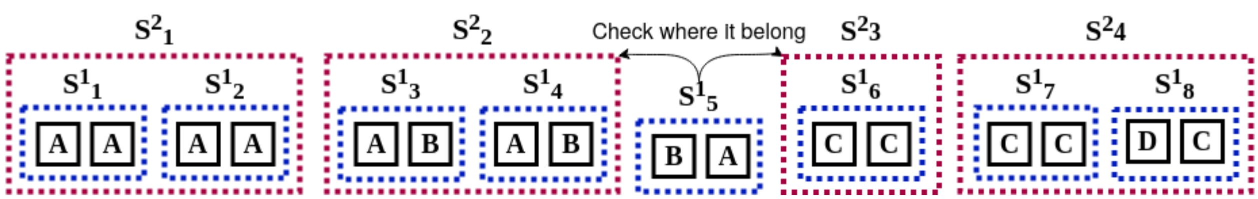

(b) Step 1



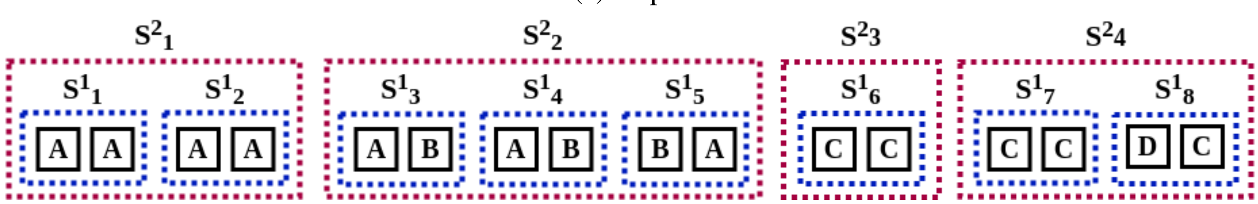

(c) Step 2



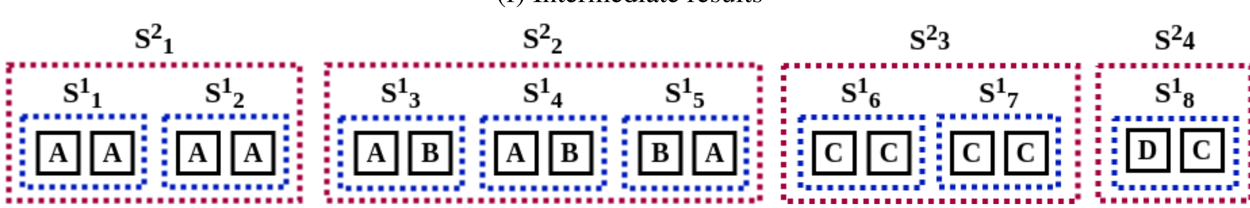

(d) Step 3




(e) Step 4




(f) Intermediate results




(g) Segments after completion of boundary adjustment at level 1

Figure 3.8: Intermediate steps for boundary adjustment at level 1 on an example.

To understand the working in more details, Fig. 3.8 shows the intermediate steps for boundary adjustment at level 1. Starting with the initialization step (Fig. 3.8a), the boundary is adjusted for the first two top level segments i.e. for $S_1^2$ and $S_2^2$. For these segments, $S_2^1$ and $S_3^1$ are the boundary sub-segments at level 1 and are thus checked for their similarity with the two parent segments. Fig. 3.8b shows that $S_2^1$ boundary segment is removed from it parent segment $S_1^2$ and is checked for similarity with $S_1^2$ and $S_2^2$ using equation 3.2. As $S_2^1$ is more similar to its parent segment, it is assigned back to $S_1^2$ as can be observed from in the next step (Fig. 3.8c). Next step is to remove $S_3^1$ from $S_2^2$ and check its similarity with $S_1^2$ with $S_2^2$. This procedure is repeated multiple times until all the boundaries are adjusted at level 1. Fig. 3.8e shows that sub-segment $S_5^1$ is checked for its belonging and is moved from $S_3^2$ to $S_2^2$. Fig. 3.8f shows the segments after the adjustment of $S_5^1$. Fig. 3.8g shows the segments after the boundaries are adjusted at level 1.

### 3.2.2 Final merging of segments

After adjusting the segment boundaries, the neighboring segments having similar location distribution are merged together. This can be understood from Fig. 3.7.d where segments $S_3^2$ and $S_4^2$ are merged together. I can also be observed in Fig. 3.6 where segments 3 and 4 as well as segments 5 and 6 in step 2 (3.6.d) are merged to give segments 3 and 4 in step 3 (3.6.e). The segments are merged if the cosine similarity between the location histograms of the two segments is above a threshold value. The cosine similarity between two histograms $h_1^{loc}$ and $h_2^{loc}$ is given by the following equation:

$$S_{cos}(h1, h2) = \frac{\sum_{j \in \mathcal{L}} h_1^{loc}[j] h_2^{loc}[j]}{\sqrt{\sum_{j \in \mathcal{L}} \left(h_1^{loc}[j]\right)^2} \cdot \sqrt{\sum_{j \in \mathcal{L}} \left(h_2^{loc}[j]\right)^2}} \tag{3.4}$$

This process is also outline in Algorithm 1. This approach is robust to presence of noise in between the activities (e.g., waking up from night sleep, having water or going to bathroom, and sleeping again). Each resulting segment is thus fairly homogeneous. These segments can be considered as approximations of unlabeled activities and will be used to assess routine uniformity in further steps.

**Algorithm 1** Modified SEEDS Algorithm

**Input:** Location sequence, Number of levels ($\mathbb{L}$), Thresholds ($\epsilon_1, \epsilon_2$)

**Output:** Contiguous homogeneous segments of locations

Initialize the hill structure as shown in Fig. 3.7

▷ Segment boundary adjustment from top level to bottom level

**for** $\ell$ from $\mathbb{L}$ to 1 **do** ▷ $\mathbb{L}$: Top level

  **for** k from 1 to $|S^{\mathbb{L}}| - 1$ **do**

    $S_a^l \leftarrow$ boundary sub-segment of $S_k^{\mathbb{L}}$ at level $\ell$

    $S_b^l \leftarrow$ boundary sub-segment of $S_{k+1}^{\mathbb{L}}$ at level $\ell$

    **if** $HI(h^{loc}_{S^{\mathbb{L}}_{k+1}}, h^{loc}_{S^l_a}) - HI(h^{loc}_{S^{\mathbb{L}}_k \setminus S^l_a}, h^{loc}_{S^l_a}) > \epsilon_1$ **then**

      Move $S_a^l$ to $S_{k+1}^{\mathbb{L}}$

    **else if** $HI(h^{loc}_{S^{\mathbb{L}}_k}, h^{loc}_{S^l_b}) - HI(h^{loc}_{S^{\mathbb{L}}_{k+1} \setminus S^l_b}, h^{loc}_{S^l_b}) > \epsilon_1$ **then**

      Move $S_b^l$ to $S_k^{\mathbb{L}}$

▷ Top level adjacent segment merging

**for** k from 1 to $|S^{\mathbb{L}}| - 1$ **do**

  **if** $S_{cos}(S_k^{\mathbb{L}}, S_{k+1}^{\mathbb{L}}) > \epsilon_2$ **then**

    Combine $S_k^{\mathbb{L}}$ and $S_{k+1}^{\mathbb{L}}$ into one segment

We evaluate the SEEDS generated segments by using intersection over union (IoU) measure between the generated segments and the ground truth activity segments for the synthetic dataset. The IoU is the ratio between the length of intersection of the two segments to the length of union of the two segments. Fig. 3.9 demonstrates the calculation of IoU on an example sequence. The first SEEDS generated segment completely aligns with the first ground truth segment and hence the IoU between the two segments is 1. For a ground truth segment captured in multiple segments (e.g. $S_4^{gt}$ segment in Fig. 3.9), the IoU is calculated with the segment of maximum overlap. The IoU for $S_4^{gt}$ segment is calculated with $S_4$ generated segment as 5/9 or 0.56.

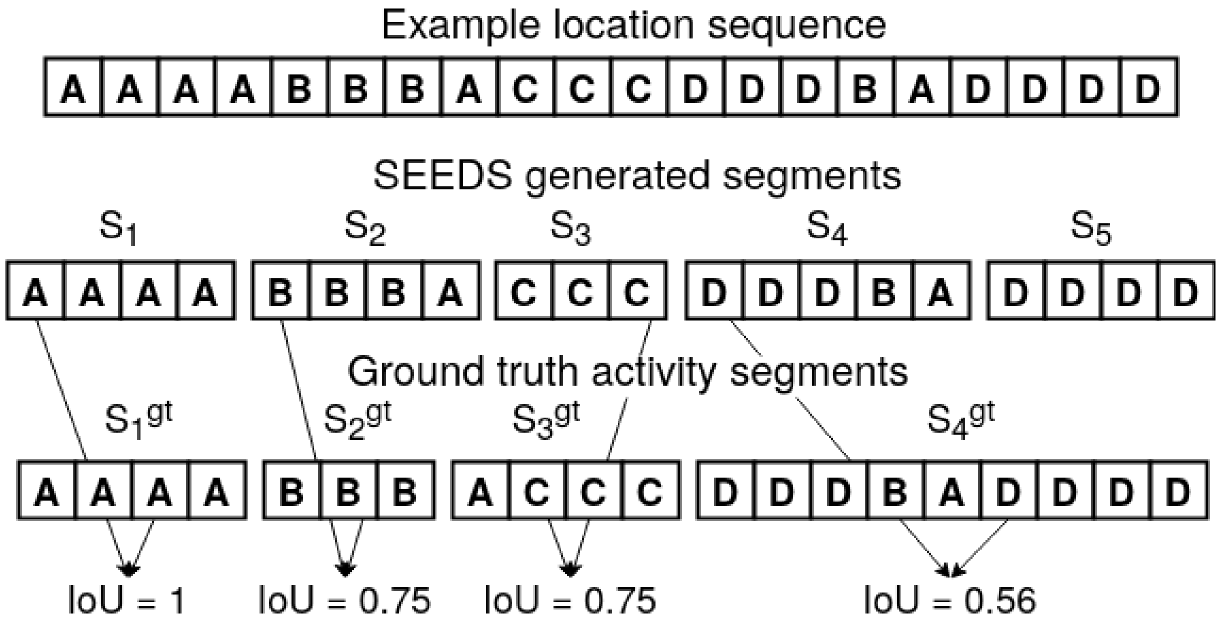



Figure 3.9: Intersection over union (IoU) between SEEDS generated segments and ground truth segments.

### 3.3 Feature representation of activities

Each activity segment resulting from the SEEDS algorithm (Section 3.2) is represented through a set of features, which describe the start and end time, duration and location distribution of the activity, as well as the location distribution of previous activities in order to introduce time-dependencies in our feature set.

Let $B$ be the number of samples of the location-based time-series collected in a day and $L$ the total number of rooms in the place where the data is collected. Also let $a$ be the $a^{th}$ activity occurring during the day. This will be described by three feature vectors, which represent its start and end time $\mathbf{h}_{\mathcal{C}}^{\mathbf{S}} \in \mathbb{R}^B$, duration $\mathbf{h}_{\mathcal{C}}^{\mathbf{D}} \in \mathbb{R}^B$, location distribution $\mathbf{h}_{\mathcal{C}}^{\mathbf{LD}} \in \mathbb{R}^L$, and location of previous activities $\mathbf{h}_{\mathcal{C}}^{\mathbf{PA}} \in \mathbb{R}^L$.

The activity time vector $\mathbf{h}_{\mathbf{a}}^{\mathbf{S}} \in \mathbb{R}^B$ for the $a^{th}$ activity will have all its elements zero, except the elements that are located between the start time $s_a$ and end time $e_a$ of the corresponding activity, which will be one:

$$h_a^S[j] = \begin{cases} 1, & s_a \leq j \leq e_a \\ 0, & otherwise \end{cases}, \quad j \in [1, B] \tag{3.5}$$

Similarly, if $d_a$ is the duration of the $a^{th}$ activity, then the corresponding duration vector $\mathbf{h_a^D} \in \mathbb{R}^B$ will be defined as follows:

$$h_a^D[j] = \begin{cases} 1, & 1 \leq j \leq d_a \\ 0, & otherwise \end{cases}, \quad j \in [1, B] \tag{3.6}$$

The location vector $\mathbf{h_a^{LD}} \in \mathbb{R}^L$ for the $a^{th}$ activity represents the location distribution of the activity. The $j^{th}$ element of the vector $\overrightarrow{h}_a^{LD}[j]$ will include the number of times room $j$ was visited during the $a^{th}$ activity. Similarly the previous activity vector for the $a^{th}$ activity, denoted as $\mathbf{h_a^{PA}} \in \mathbb{R}^L$, represents the activities that occurred in each room prior to activity $a$ on the same day, where $L$ is the total number of rooms. The $j^{th}$ element of the previous activity vector $\overrightarrow{h}_a^{PA}[j]$ will include the number of times room $j$ was visited prior to the $a^{th}$ activity at the same day.

If we have a set of activities $\mathcal{C}$, then we can represent their total start and end time vector $\mathbf{h_{\mathcal{C}}^S} \in \mathbb{R}^B$, duration vector $\mathbf{h_{\mathcal{C}}^D} \in \mathbb{R}^B$, location vector $\mathbf{h_{\mathcal{C}}^{LD}} \in \mathbb{R}^L$, and previous activities vector $\mathbf{h_{\mathcal{C}}^{PA}} \in \mathbb{R}^L$ as the sum of each of the corresponding vectors for each of the activities that are included in set $\mathcal{C}$:

$$\mathbf{h_{\mathcal{C}}^S} = \sum_{\mathrm{a} \in \mathcal{C}} \mathbf{h_a^S} \tag{3.7}$$

$$\mathbf{h_{\mathcal{C}}^D} = \sum_{\mathrm{a} \in \mathcal{C}} \mathbf{h_a^D} \tag{3.8}$$

$$\mathbf{h_{\mathcal{C}}^{LD}} = \sum_{\mathrm{a} \in \mathcal{C}} \mathbf{h_a^{LD}} \tag{3.9}$$

$$\mathbf{h_{\mathcal{C}}^{PA}} = \sum_{\mathrm{a} \in \mathcal{C}} \mathbf{h_a^{PA}} \tag{3.10}$$

### 3.4 Hierarchical clustering

This section answers our second question on how to find similar segments repeating in a similar fashion. The activity segments obtained through SEEDS algorithm (Section 3.2) are represented through four features as detailed in Section 3.3. The segments which are similar in location distribution, time, and duration, and also have a similar set of previous activities are grouped together to assess the uniformity of routine patterns. The segments are grouped together through a hierarchical graph-based region growth algorithm based on the approach proposed in [39].

The daily routine data for multiple days is represented as a graph $G(V, E)$ where vertex $V$ denotes a set of activity segments or a clusters of similar activity segments connected by edges $E$ with edge weights $w$. The edge weight $w$ between two connected vertices is the similarity score between the two activity segments or clusters of activities. Initially, all segments are considered as unique vertex with no connections or edges. The segments that are similar in terms of the location distribution, start time, duration and previous activities are then connect by an edge. This is repeated until no more changes in the graph can be made. The final connected components of the graph represent the clusters of similar activity segments, which will then be used to assess the uniformity of the daily routine patterns.

Algorithm 2 summarizes the steps of the proposed hierarchical clustering approach. The algorithm takes a list of activity segments or clusters ($\mathcal{L}$) and three threshold values ($\epsilon_1, \epsilon_2, \epsilon_3$) as an input, and outputs the list of clustered segments ($\mathcal{C}$). For an activity segment $a$, the variable $C(a)$ denote the connected component or cluster of $a$. Clustering of segments is performed in two separate phases with different similarity criteria. At the first phase, the grouping of vertices (i.e., clustering of activities) is performed by looking at the similarity between activity clusters with respect to their start/end time and duration, according to:

$$SS_1(C(n), C(m)) = S_{cos}(h_n^S, h_m^S) * S_{cos}(h_n^D, h_m^D) \tag{3.11}$$

At the second phase, similarity is evaluated with respect to previous activity vector space, as follows:

**Algorithm 2** Hierarchical Clustering Algorithm

**Input:** List $\mathcal{L}$ of activities, Thresholds $(\epsilon_1, \epsilon_2, \epsilon_3)$
**Output:** List $\mathcal{C}$ of clusters

Create a graph $G(V, E, w)$ of activities
Initialize $V$ with $\mathcal{L}$ ▷ assign each activity to a vertex
**repeat**
  **for** $\forall\, n, m \in V$ **do** ▷ for all pairs of vertices
    Compute $w(n, m) = SS_1(n, m)$ ▷ equation 3.11
    **if** $S_{cos}(h_n^{LD}, h_m^{LD}) \geq \epsilon_1$ & $w(n, m) \geq \epsilon_2$ **then**
      Connect $n$ and $m$ to same vertex
**until** no change in $G$
**repeat**
  **for** $\forall\, n, m \in V$ **do** ▷ for all pairs of vertices
    Compute $w(n, m) = SS_2(n, m)$ ▷ equation 3.12
    **if** $S_{cos}(h_n^{LD}, h_m^{LD}) \geq \epsilon_1$ & $w(n, m) \geq \epsilon_3$ **then**
      Connect $n$ and $m$ to same vertex
**until** no change in $G$
Assign vertices $V$ to list of clusters $\mathcal{C}$

$$SS_2(C(n), C(m)) = (1 - Dif(s)) * S_{cos}(h_n^{PA}, h_m^{PA}) * S_{cos}(h_n^{D}, h_m^{D}) \tag{3.12}$$

where Dif(s) is the absolute difference between the average start times of the activities of two cluster.

## 3.5 Routine Assessment

The daily routines are assessed by evaluating the clusters of similar activities resulting by the hierarchical clustering algorithm (Section 3.4). To evaluate the activity clusters we have used three measures: number of clusters, root mean square error (RMSE), and mean absolute error (MAE).

The final number of clusters is an important measure to assess the uniformity of routines as it captures the consistency in the daily routines. Lower number of clusters tend to be more indicative of more uniform routines, in which more activities are performed on a routine basis, therefore more activities get clustered together resulting in lower number of clusters. For a less uniform ADL pattern only a few activities are performed in routine. Therefore, only a few activities are grouped together increasing the resulting number of clusters.

A cluster of similar activities signify an activity performed on multiple days at a similar time or after a similar order of activities and having a similar location distribution. Even though a cluster groups similar activities, these activities show variations with respect to start time and duration. The low uniformity routines show a higher value of these variations. We calculate these variations using RMSE and MAE scores. For an ADL routine resulting in $N$ clusters and if $s_i$ and $d_i$ denote the start time and duration of the i$^{\text{th}}$ activity (a), $\mu_C^s$ and $\mu_C^d$ denote the mean start time and mean duration of activities of cluster C having n(C) number of activities, then the RMSE and MAE scores are calculated by summing the measures over each cluster as given in equations below:

$$RMSE = \sum_{C=1}^{N} \frac{1}{\sqrt{n(C)}} \left( \sqrt{\sum_{a_i \in C} (s_i - \mu_C^s)^2} + \sqrt{\sum_{a_i \in C} (d_i - \mu_C^d)^2} \right) \tag{3.13}$$

$$MAE = \sum_{C=1}^{N} \frac{1}{n(C)} \left( \sum_{a_i \in C} |s_i - \mu_C^s| + \sum_{a_i \in C} |d_i - \mu_C^d| \right) \tag{3.14}$$

We present the results of our approach on the proposed synthetic data as well as the collected data for 4 subjects. We first present the results for SEEDS generated segments on the three types of synthetic routines by comparing them with the ground truth segments using intersection over union (IoU) measure as explained in Section 3.2. We then present the findings of hierarchical clustering on the two types of synthetic data and the collected data with respect to the measures discussed in Section 3.5. We also compare our approach with two the baseline approaches which uses sequence matching and Poincare plots for routine assessment.

Table 4.1 provides the IoU values for the sequences in the different types of synthetic routines. A higher IoU value depict a better overlap between the generated activity segments and the ground truth activity segments. Type 1 synthetic routines are the most uniform routine sequences with no noise activities. The SEEDS algorithm achieves 90-95% accuracy for Type 1 routines. We found that the SEEDS accuracy on Type 1 reduces with increasing value of percentage standard deviation ($X$). A possible explanation to this is that with higher standard deviation, the duration of smaller activities (e.g. personal hygiene) might become very small. Therefore the activity gets considered as noise and is merged with neighboring activity segments. From the table we

| **%SD** (x) | **Type 1** | **Type 2** Noise count (N) | | | **Type 3** Consistency (P) | | | |
|---|---|---|---|---|---|---|---|---|
| | | 20 | 30 | 40 | 0.3 | 0.5 | 0.7 | 0.9 |
| 5 | 0.950 | 0.798 | 0.796 | 0.755 | 0.833 | 0.849 | 0.898 | 0.899 |
| 10 | 0.943 | 0.796 | 0.778 | 0.738 | 0.895 | 0.847 | 0.882 | 0.896 |
| 15 | 0.933 | 0.800 | 0.773 | 0.737 | 0.890 | 0.868 | 0.862 | 0.890 |
| 20 | 0.931 | 0.792 | 0.784 | 0.736 | 0.895 | 0.849 | 0.856 | 0.895 |
| 25 | 0.921 | 0.779 | 0.761 | 0.720 | 0.861 | 0.847 | 0.869 | 0.861 |
| 30 | 0.903 | 0.772 | 0.755 | 0.709 | 0.883 | 0.819 | 0.848 | 0.883 |

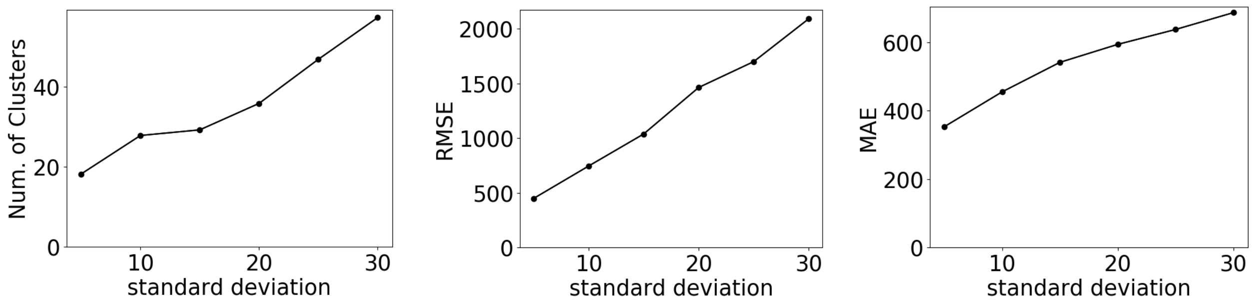
Table 4.1: Average intersection over union (IoU) values between SEEDS generated activity segments and ground truth segments.

observe that SEEDS performance on Type 2 routines is consistent over the routines with different values of $N$ where $N$ represents the number of noise activities allowed each day. This shows that SEEDS algorithm is robust to presence of noise in between activities. The results demonstrate that the SEEDS algorithm is able to capture contiguous homogeneous segments within the activity sequence for each day. The results also demonstrate that these segments are very near to ground truth segments and thus can be considered as approximations of activities within the day.

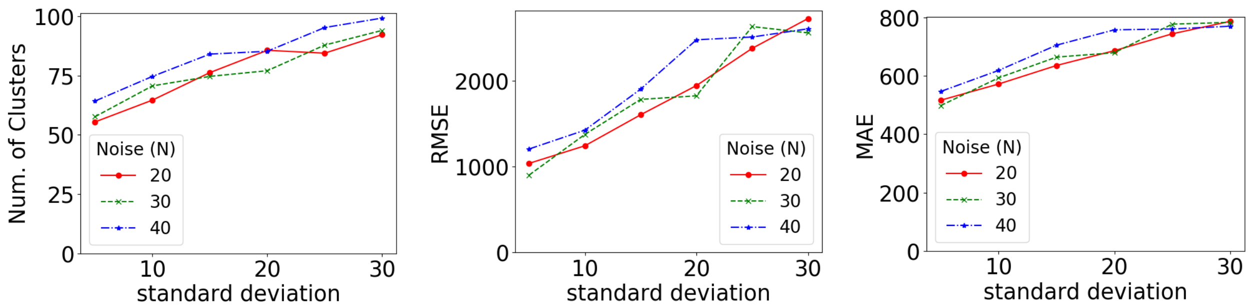
We present a detailed analysis of change in routine uniformity with respect to the three measures, as described in Section 3.5 for the proposed synthetic data. Fig. 4.1 presents the trends in uniformity of routines in the different types of synthetic data. We found that the number of clusters increases with the decrease in the uniformity of routines. As the routine uniformity decreases, lesser number of activities are repeated in similar fashion. This leads to lesser number of similar activities resulting in higher number of activity clusters. From the figures we can observe that Type 1 routines have lowest number of resulting clusters being the more uniform routines. Type 2 routines incorporate noise activities which are often not a part of routines. Presence of these activities thus increases the number of resulting clusters. For Type 3 routines we found an increase in number of clusters with decreasing $P$ values (increasing inconsistencies). Type 3 routines with P = 0.3 and 0.5 being the least uniform routines results in maximum number of resulting clusters.

The RMSE and MAE scores signify the variations within the similar activity segments of a cluster with respect to start time and duration. As the randomness in a routine increases, the activity's start time as well duration shows more variations thus increases the RMSE and MAE scores of the routine. Fig. 4.1 shows the trends of the two measures for the three types of synthetic routines. The results for Type 1 and Type 2 present a clear increase in both scores with increasing standard deviation. For Type 3, we observe an increase and then a decrease in the trends of RMSE and MAE scores. A possible explanation to this is the fact that with a high number of clusters, many clusters contain single segments which do not contribute to RMSE or MAE scores.

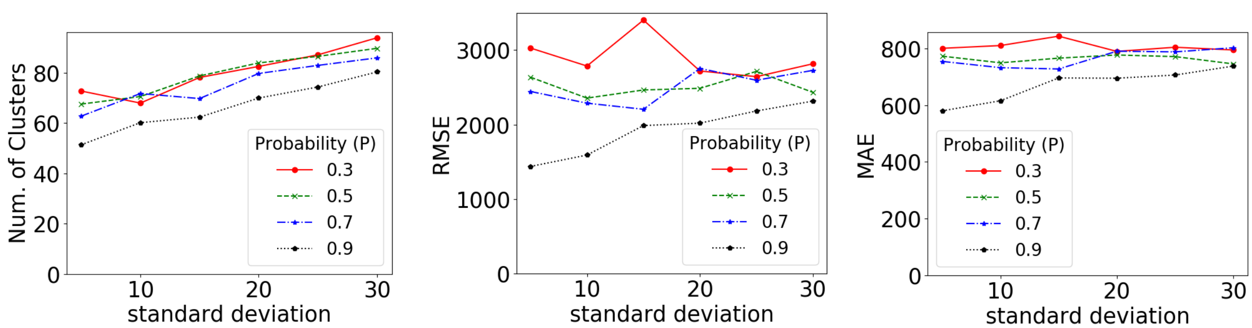
Fig. 4.2 shows the visualization for a type 2 synthetic routine before (left; raw sequences of locations) and after (right; similar activity segments) clustering. The left magnified image shows




(a) Type 1 with increasing percentage standard deviation ($X$) in activity duration.




(b) Type 2 with increasing percentage standard deviation ($X$) in activity duration and different number of noise instances $N \in \{20, 30, 40\}$.




(c) Type 3 with increasing percentage standard deviation ($X$) in activity duration and different consistency levels $P \in \{0.9, 0.7, 0.5, 0.3\}$.

Figure 4.1: Results of hierarchical region growth approach on Synthetic Data.

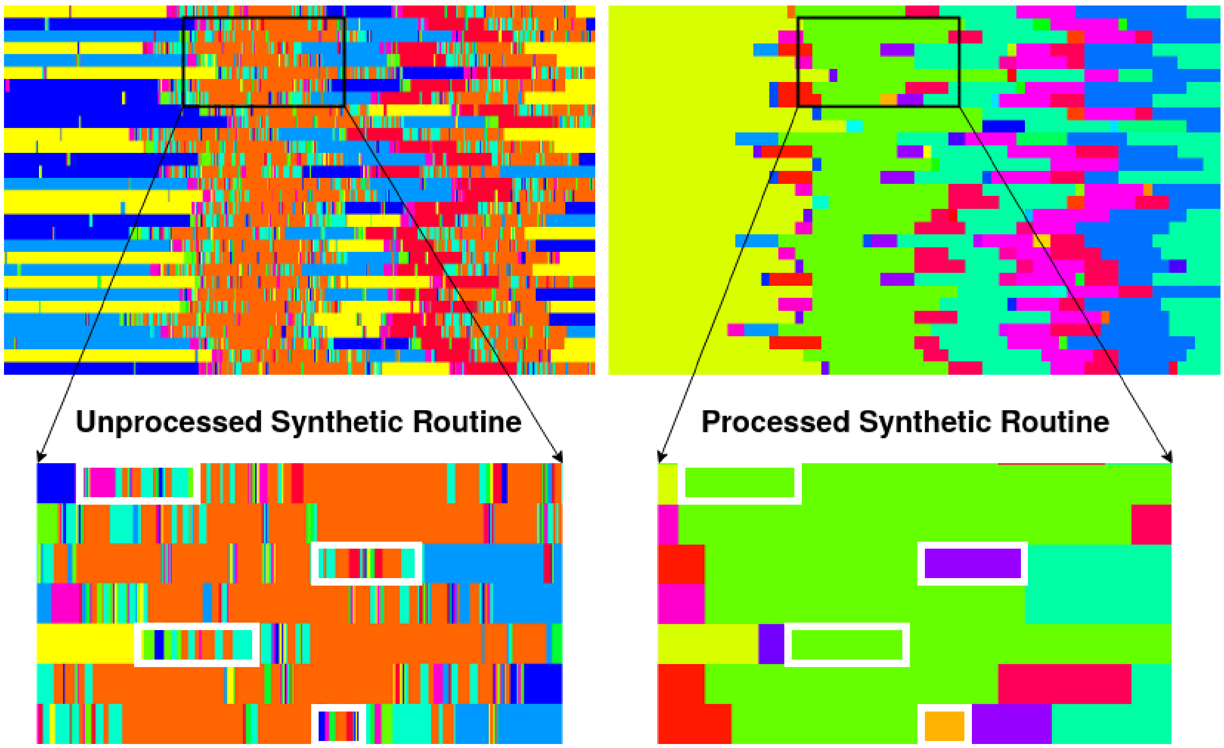
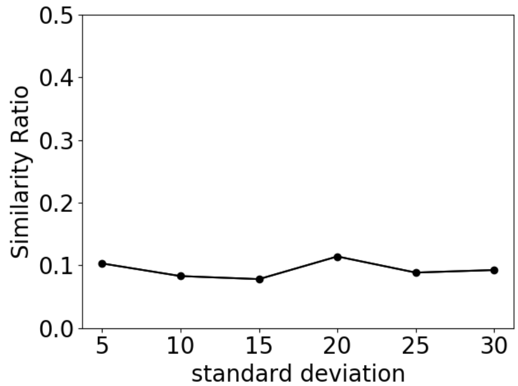


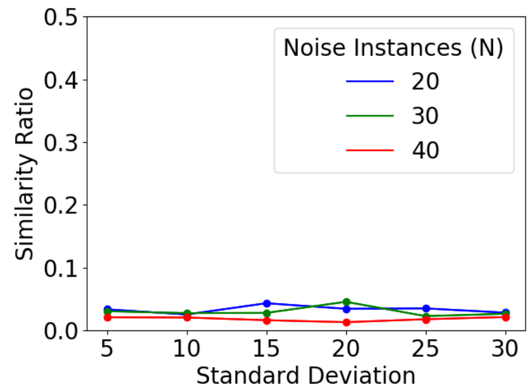

Figure 4.2: Visualization of unprocessed Type 2 synthetic routine (left; each location shown with a unique color) and processed data after applying SEEDS and hierarchical clustering (Right; similar segments shown with unique colors).

entertainment activity for six days with large amounts of noise and variations in the start time and duration of the activity. From the right magnified image, it can be observed that SEEDS was able to bundle these noise into segments and the hierarchical clustering algorithm combined these segments into a single cluster . The right magnified image also highlights some false segments formed by SEEDS (e.g. the segments highlighted by white rectangles in the image) as well as errors at the boundary of the segments.

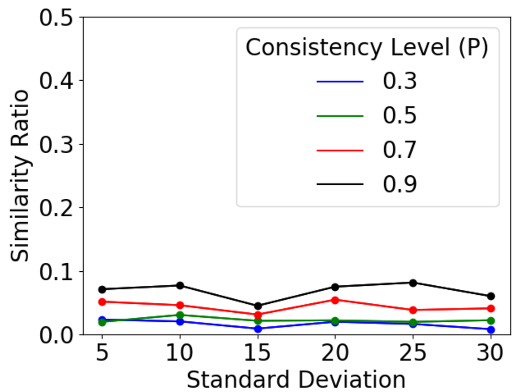
We compare the performance of our approach with two baseline methods. The method explained in [24] uses Gestalt Sequence Matching to compute similarity between the routine sequences of two days. Gestalt Sequence Matching (GSM) uses the concept of longest common sub-sequence to calculate the similarity between two sequences [31]. This method compares the routine sequences of consecutive days to look for similarity between the routines. The similarity is computed using equation 1.1 as explained under related works. As same activities on different




(a) Type 1




(b) Type 2

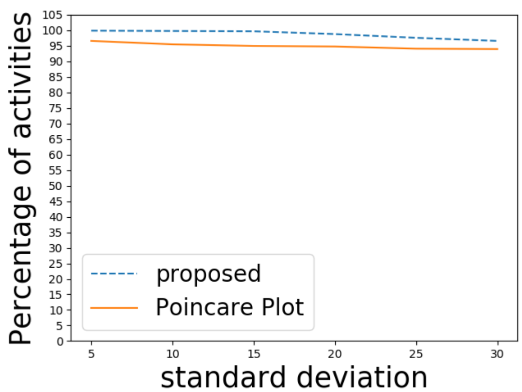



(c) Type 3

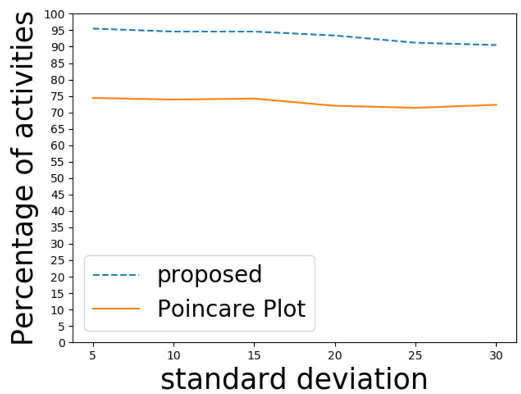
Figure 4.3: Gestalt sequence matching results on the three classes of synthetic data.

days might have huge variations in terms of start and end time and the feature distribution might not have an exact match, the longest common sub-sequence approach is not able to capture the similarity between the same activities. Also, the GSM approach is highly affected by the presence of non routine activities or noise stemming from various factors such as multitasking (example, cooking and cleaning can be done simultaneously) or temporary interruption in an ongoing activity (example:waking up for toileting during a long sleep), which are inevitable part of human routine. Fig. 4.3 provides the results of GSM on our synthetic data. It can be observed that the method is not able to differentiate between the high uniform (SD = 5, 10, 15) and less uniform (SD = 20, 25, 30) routines of the three types of synthetic data. In Type 3 routines the difference across the consistency levels (P) is negligible. On the other hand, our approach is able to captures variations in similarly occurring activities and inconsistencies in the daily routines and is robust to the presence of non routine activities (as depicted by Type 2 routines).

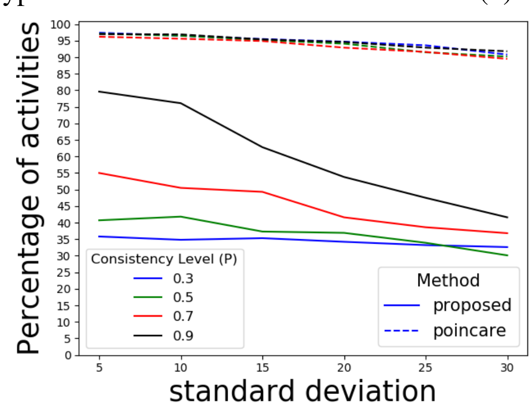
Another baseline we compare our approach with uses Poincare plots (PP) as proposed in [1]. This method first recognises a few key activities using a rule based in-house developed CAR-classifier [29] followed by routine assessment using Poincare Plots. This method makes an assumption that certain activities are performed at a specific time and for a specific duration. Once the activities are recognised, same activities occurring at a similar time on two consecutive days are paired and plotted on a PP. The plotted points are then assessed by fitting an ellipse around the points. The method differentiates between more uniform routines and less uniform routines using the centroid and the two axis of the fitted ellipses. Even though Poincare plots provide good information regarding the repetition of activities, the approach suffers because of the inconsistencies in




(a) Type 1




(b) Type 2

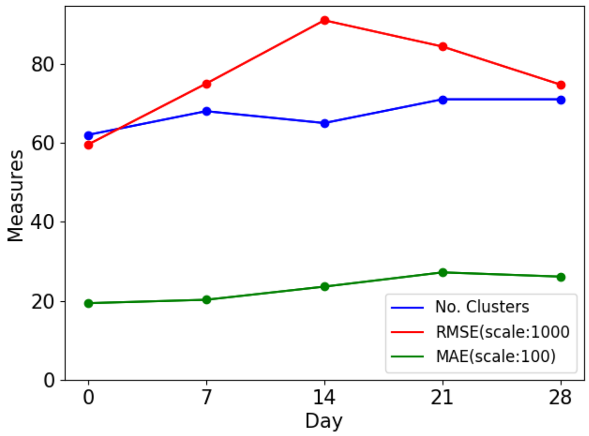



(c) Type 3

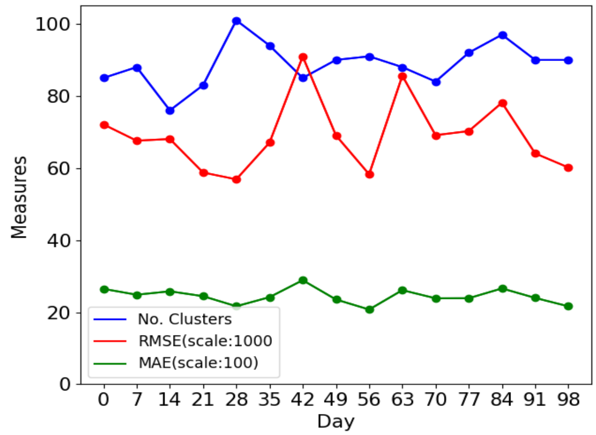
Figure 4.4: Percentage of total activity paired by Poincare plot approach [1] VS percentage of total activities clustered by our proposed clustering algorithm.

the daily routines. As routine activities repeat with different frequencies and not all activities form a part of routines, it becomes difficult to find pairs of similar activities in inconsistent routines and might require human efforts. Fig. 4.4 compares the percentage of total activities that were paired for Poincare plot approach and the percentage of total activities that were clustered by our proposed hierarchical clustering approach on the synthetic data. From these figures it can be observed that Poincare plot approach performed very well on Type 1 routines. This is because Type 1 routines follow activities in a defined order and thus the activity schedule is same for every day. For Type 3 routines, we observe a considerably low number of activities pairs as the routines are inconsistent and for many activities, there is no matching activity on the previous or next day. On the other hand, our approach was able to cluster significantly good amount of activities for both consistent and inconsistent routines.

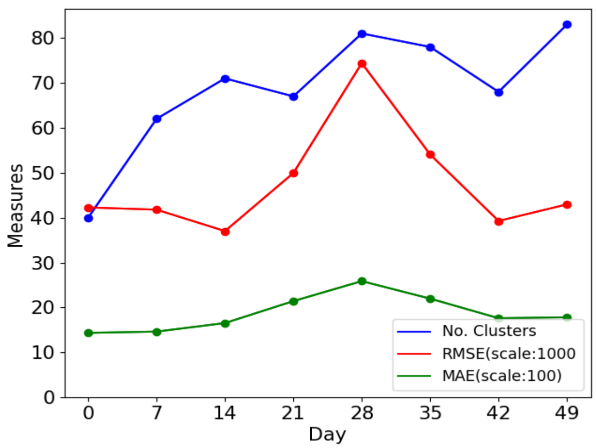
Fig. 4.5 provide the evaluation of the routine data collected from the four subjects. For all subjects, routine evaluation is performed on a window of 14 consecutive days and a step size of 7 days. Therefore, for subject 1, the routine analysis is done from day 1 to 14, day 7 to 21, day 14 to 28 and so on. The X-axis of graphs in Fig. 4.5 list the start day of every analysis period. As described earlier, subject 1 pays frequent long visits to friends during mid day while subject 3 frequently stays with friends overnight and sometimes for the whole day. Therefore, the number of activities captured within the house for both the subjects is less. This can be related to a lesser number of room transitions every day for the two subjects (Subject 1: 50 & Subject 3: 52). Because of the lower number of captured activities in case of subject 1 and 3, the variations captured in daily routines are also less thus showing a uniform routine for the two subjects as can be observed in the graphs. During the 3$^{rd}$ and 4$^{th}$ week of data collection (i.e. from day 21 to 35 in Subject 3's graph), Subject 3 mostly stayed at home and therefore more activity data was captured during this period. This resulted in an increased RMSE and MAE scores during this period as compared to days when the subject was not at home. Subject 2 is a healthy and active person who mostly stays at home except for small visits outside house like evening walks and has a high number of room transitions each day (avg. 97). From the three measures of subject 2, it can be observed that subject 2 has



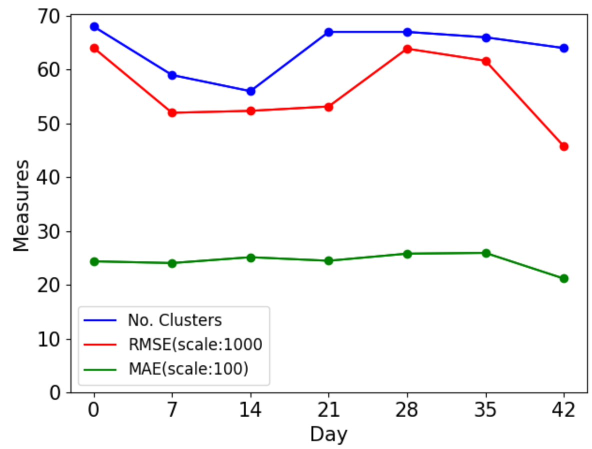

(a) Subject 1




(b) Subject 2




(c) Subject 3




(d) Subject 4

Figure 4.5: Routine uniformity trends for real-life data from four subjects.

a more uniform routine for some weeks while a lesser uniform routine on other weeks. Subject 4 has been diagnosed of Parkinson's disease and mostly stays at home. Parkinson's disease is a progressive nervous system disorder that affects movement. Therefore, Subject 4 has a lower number of average room transitions (avg. 74) and has a stable routine which can be observed from the three relatively stable measures for Subject 4.

# 5. DISCUSSION

This research presents a unique approach towards the longitudinal assessment of daily routines of elderly people in order to identify the possibility of development of chronic diseases such as dementia or depression. These diseases are of a slow progressive nature and the symptoms are often difficult to identify during the early stages. Therefore a long term monitoring of activities can provide insightful information on the development and progression of such diseases. We use a network of non-intrusive motion sensors installed at various locations within the individual's house to continuous monitor in-house activities of the subjects. Our approach process sequences of locations obtained from these sensors to assess the routine uniformity. Our proposed hierarchical clustering algorithm is able to identify potential activity patterns that contribute towards the routine of the person as well as captured the variations in the repetition of these patterns.

In our proposed approach, we represent the time-series sensor data as images (or 2D matrix) which allows us a flexibility to use a wide range of image processing algorithms. The raw data for each day is filtered using median filtering and then divided into homogeneous segments using SEEDS algorithm. As specific activities are often performed in specific rooms, each activity is associated with a particular distribution of rooms. Therefore, the segments resulting from SEEDS algorithm can be treated as approximations of activities. The activity segments for multiple days are then clustered together using a graph based hierarchical clustering algorithm. The hierarchical clustering algorithm clusters together the activity segments which are similar in location distribution, start time, duration or have a similar order of previous activities. By doing this, our proposed approach has advantage over the existing baselines as we are able to capture activity patterns repeating daily or infrequently as well as the variations within the repetition of same activity (e.g. having breakfast at different times is a more varying activity than always having breakfast at 8:00 am). The use of SEEDS algorithm makes our approach robust to noise activities. Another advantage of our approach is that the uniformity measures are not affected by a sudden change in the daily routine of the person. If a person follows two different sets of routines (e.g. daily routine followed on week 1, 2 and 5 is same, routine for week 3, 4, and 6 is same but the two routines are different) or changes the routine abruptly (e.g. used to sleep from 11:00pm to 6:00 am but changed the sleeping hours to 6:00 am to 12:00 pm), they will still have a similar RMSE and MAE values if the individual sets of routines are followed uniformly.

To assess the real-life data, we used a sliding window approach with window size of 14 days and strides of 7 days. These numbers are selected because of the small size of available data-set. Cognitive diseases are slow progressive diseases with a very gradual decline in daily routines. Therefore, in real-life monitoring, use of a larger window size of around 3 to 4 months with strides of 1 month would give more precise results. Also, as daily routines may not have similar uniformity during two successive assessment period. This can be seen in the routine data of subject 2 in Figure 4.5.b where the uniformity for some weeks is higher than others. Even though subject 2 has many instances of low or high uniformity, the average variations during the whole period on 100 days do not vary much. Similarly, a long term assessment (e.g. 2 to 5 years), would have multiple instances of high and low variability routines but a significant increase in the average variations during the period might indicate the development or progression of cognitive diseases. Our approach validates the uniformity changes within the daily routines of individual subject. The three measures used to tracks the uniformity trends are influenced by the living environment, number of activities performed and the amount of activity data captured for each individual subject. Therefore a comparison between different subjects may not be valid.

# 6. SUMMARY, CONCLUSION AND FUTURE EXTENSIONS

## 6.1 Summary

In this thesis we present a novel approach towards the longitudinal assessment of daily routines of elderly people in order to identify the possibility of development of chronic diseases such as dementia or depression. To achieve this, we use non-intrusive in-house activity monitoring system and propose a method to capture the gradual changes in the uniformity of daily routines. For long-term in-home monitoring, we use a network of non-intrusive motion sensors installed at various locations within the individual's house to record continuous in-house activity data from four subjects. These sensors provide the motion coordinates within the house which are mapped to the specific room/locations of the house. To reduce sensor noise in the recorded data, we apply median filtering on the sequence of mapped locations thus reducing the sequence length from 17,280 (number of samples recorded each day at 0.2Hz) to 1440 (number of samples per minute each day). As data collection is a resource-intensive process and it is difficult to obtain fully labelled data in real-life, we propose a synthetic data generated for the purposes of this problem. The synthetic routines are generated but introducing systematically controlled variations in a predefined activity routine.

Each day's sequence is then divided into multiple contiguous homogeneous segments. Since each activity can be characterized by a particular distribution of locations, these segments can be considered as approximations of daily activities. To divide the sequences into homogeneous segments, me propose a modification in a widely used image segmentation algorithm called Superpixels Extracted via Energy-Driven Sampling (SEEDS). The segments generated from SEEDS algorithm over a span of several successive days are than grouped together using a graph based hierarchical clustering algorithm. The clustering algorithm groups segment which are similar in terms of location distribution, time and duration of occurrence, and previous set of activities. Each of the resulting cluster of similar activities represent activity patterns that repeat in a similar fashion over multiple days.

The assessment of routine uniformity if done by evaluating the clusters of similar segments generated by the hierarchical clustering algorithm. These clusters are evaluated through RMSE and MAE scores as well as the number of resulting clusters. We present the evaluation results of our proposed approach on different types of synthetic routines as well as on the real-life data collected from the four subjects. We also present results to compare our approach with two baseline methods which use sequence matching and Poincare plots to assess daily activity routines.

## 6.2 Conclusion

We proposed an approach to assess the daily routines of elderly people in a smart home environment in order to capture gradual changes in the routine uniformity. To evaluate our approach, we collected continuous activity data from four subjects as well as proposed a synthetic dataset to capture the various variations in daily routines. The results show that the modified SEEDS algorithm was able to divide the time-series data into homogeneous segments and these segments can be considered as the approximations of real activities. Our proposed hierarchical clustering algorithm was able to find similar activity segments repeating on several days. These clusters of similar activity segments represents the activities which form the routine of the person. The evaluation results shows that our approach is able to capture differences in the uniformity of routines in different types of synthetic routines. The results for real-life data from four subjects demonstrate that the sliding window approach for real-life routine assessment has potential to capture gradual changes in routine uniformity over a long period of time and can provide insightful information towards the development and progression of cognitive diseases. The comparison of our approach with the two baseline methods highlights the advantages of our proposed method of routine assessment. Our approach is robust to the presence of noise activities, is able to capture variations within routinely repeated activities as well the inconsistencies in the daily routines.

## 6.3 Limitations

Even though our proposed approach outperforms the baseline approaches, our approach faces some limitations. The real-life evaluation through our approach depends on how much time the person stays at home. This was seen in case of subject 1 and 3 who spend a major time outside their house. Because of this, the number of activities captured are less and the uniformity estimates may not be accurate. As the real-life data is often unlabelled, there is no formal evaluation of SEEDS generated activity segments for real-life data. We make the use of our proposed synthetic data to evaluate SEEDS algorithm for unsupervised activity segmentation and used the parameters of best accuracy for the real routine assessment. Another limitation of our approach is its high dependence on the hyperparameters. The parameters in our approach are the number of segments and levels during the initialization of SEEDS algorithms and the various thresholds in SEEDS as well as the proposed hierarchical clustering algorithm. These hyperparameters require diligent fine tuning which is sometimes difficult due to lack of labelled data. Another limitation of our approach is its robustness towards noise activity. Through our synthetic data we show that our approach is robust to the presence of noise but this is valid upto a certain limit. A large amount of noise activities degrades the performance of unsupervised activity segmentation through SEEDS algorithm. This is because presence of noise activities results in splitting of an activity into smaller segments or merging different activity segments into one. This can be observed in Fig. 4.2. Some of these limitations can be overcome as we shall discuss under the future works.

## 6.4 Future Extensions

For this research, we only used motion information but our approach are also be used with other types of data streams. Our approach can be extended to include more information such as appliance usage (e.g. television, refrigerator, and stove), furniture usage (e.g. sofa, bed, and doors) though pressure sensors or binary sensors, raw motion coordinates and motion intensity. Including such features would provide more insightful information about ongoing activities. From Fig. 4.2 we observe that SEEDS algorithm is vulnerable to errors at the boundary of the segments. In some instances, SEEDS algorithm merges two different activity segments into one or splits one activity segment into multiple different segments. These errors are difficult to overcome through motion data alone but having different sensor information is expected to improve SEEDS performance. Another possible improvement to our approach can be the use of more descriptive features. Our method uses previous activity information as a "bag of words" approach and does not include the order of the activities. This can be improved to include sequential information about the previous activities.

REFERENCES

[1] P. Urwyler, R. Stucki, L. Rampa, R. Müri, U. P. Mosimann, and T. Nef, "Cognitive impairment categorized in community-dwelling older adults with and without dementia using in-home sensors that recognise activities of daily living," *Scientific reports*, vol. 7, p. 42084, February 2017.

[2] Wikipedia contributors, "Gestalt pattern matching — Wikipedia, the free encyclopedia," 2020.

[3] World Health Organisation (WHO), "Dementia key-facts," 2019. `https://www.who.int/news-room/fact-sheets/detail/dementia`.

[4] A. Association, "2019 alzheimer's disease facts and figures.," *Alzheimers Dementia*, vol. 15, pp. 321–387, 2019.

[5] C. Blankevoort, M. Heuvelen, F. Boersma, H. Luning, J. de Jong, and E. Scherder, "Review of effects of physical activity on strength, balance, mobility and adl performance in elderly subjects with dementia.," *Dementia and Geriatric Cognitive Disorders*, vol. 5, no. 30, pp. 392–402, 2010.

[6] J. Ormel, F. V. Rijsdijk, M. Sullivan, E. V. Sonderen, and G. I. J. M. Kempen, "Temporal and reciprocal relationship between iadl/adl disability and depressive symptoms in late life," *Journals of Gerontology: Series B*, vol. 57, no. 4, pp. P338–P347, 2002.

[7] H. H. Feldman, C. Jacova, A. Robillard, A. Garcia, T. Chow, M. Borrie, H. M. Schipper, M. Blair, A. Kertesz, and H. Chertkow, "Diagnosis and treatment of dementia: 2. diagnosis," *CMAJ*, vol. 178, no. 7, pp. 825–836, 2008.

[8] F. Pasquier, "Early diagnosis of dementia: Neuropsychology," *Journal of neurology*, vol. 246, pp. 6–15, 02 1999.

[9] A. Ströhle, "Physical activity, exercise, depression and anxiety disorders," *Journal of neural transmission*, vol. 116, pp. 777–784, 2009.

[10] C. M. Giebel, C. Sutcliffe, M. Stolt, S. Karlsson, A. Renom-Guiteras, M. Soto, H. Verbeek, A. Zabalegui, and D. Challis, "Deterioration of basic activities of daily living and their impact on quality of life across different cognitive stages of dementia: a european study," *International Psychogeriatrics*, vol. 26, no. 8, p. 1283–1293, 2014.

[11] G. McKhann, D. Drachman, M. Folstein, R. Katzman, D. Price, and E. Stadlan, "Clinical diagnosis of alzheimer's disease: Report of the nincds-adrda work group under the auspices of department of health and human services task force on alzheimer's disease," *Neurology*, vol. 34, no. 7, pp. 939–944, 1984.

[12] H. Dodge, T. Kadowaki, T. Hayakawa, M. Yamakawa, A. Sekikawa, and H. Ueshima, "Cognitive impairment as a strong predictor of incident disability in specific adl-iadl tasks among community-dwelling elders: The azuchi study," *The Gerontologist*, vol. 45, pp. 222–30, 05 2005.

[13] A. K. Desai, G. T. Grossberg, and D. N. Sheth, "Activities of daily living in patients with dementia: Clinical relevance, methods of assessment and effects of treatment," *CNS Drugs,*, vol. 18, p. 853–875, 2004.

[14] The Staff of the Benjamin Rose Hospital, "Multidisciplinary studies of illness in aged persons: Ii. a new classification of functional status in activities of daily living," *Journal of Chronic Diseases"*, vol. 9, no. 1, pp. 55 – 62, 1959.

[15] S. KATZ, "Assessing self-maintenance: Activities of daily living, mobility, and instrumental activities of daily living," *Journal of the American Geriatrics Society*, vol. 31, no. 12, pp. 721–727, 1983.

[16] G. Carlsson, M. Haak, C. Nygren, and S. Iwarsson, "Self-reported versus professionally assessed functional limitations in community-dwelling very old individuals," *International Journal of Rehabilitation Research*, vol. 35, p. 299–304, 2012.

[17] K. Hänsel, N. Wilde, H. Haddadi, and A. Alomainy, "Challenges with current wearable technology in monitoring health data and providing positive behavioural support," in *Proceedings of the 5th EAI International Conference on Wireless Mobile Communication and Healthcare*, p. 158–161, 2015.

[18] S. Majumder and M. J. Deen, "Smartphone sensors for health monitoring and diagnosis," *Sensors*, vol. 19, no. 9, 2019.

[19] S.-R. Ke, H. L. U. Thuc, Y.-J. Lee, J.-N. Hwang, J.-H. Yoo, and K.-H. Choi, "A review on video-based human activity recognition," *Computers*, vol. 2, no. 2, pp. 88–131, 2013.

[20] K. K. B. Peetoom, M. A. S. Lexis, M. Joore, C. D. Dirksen, and L. P. D. Witte, "Literature review on monitoring technologies and their outcomes in independently living elderly people," *Disability and Rehabilitation: Assistive Technology*, vol. 10, no. 4, pp. 271–294, 2015.

[21] B. E. Lyons, D. Austin, A. Seelye, J. Petersen, J. Yeargers, T. Riley, N. Sharma, N. Mattek, K. Wild, H. Dodge, and J. A. Kaye, "Pervasive computing technologies to continuously assess alzheimer's disease progression and intervention efficacy," *Frontiers in Aging Neuroscience*, vol. 7, p. 102, 2015.

[22] F. J. Ordóñez, P. de Toledo, and A. Sanchis, "Activity recognition using hybrid generative/discriminative models on home environments using binary sensors," *Sensors (Basel, Switzerland)*, vol. 13, no. 5, p. 5460—5477, 2013.

[23] S. Sharma and A. Ghose, "Unobtrusive and pervasive monitoring of geriatric subjects for early screening of mild cognitive impairment," in *2018 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*, pp. 179–184, March 2018.

[24] A. Alberdi, A. Weakley, M. Schmitter-Edgecombe, D. J. Cook, A. Aztiria, A. Basarab, and M. Barrenechea, "Smart home-based prediction of multidomain symptoms related to alzheimer's disease," *IEEE Journal of Biomedical and Health Informatics*, vol. 22, pp. 1720–1731, Nov 2018.

[25] H. Pirsiavash and D. Ramanan, “Detecting activities of daily living in first-person camera views,” in *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pp. 2847–2854, June 2012.

[26] T. Nguyen, J.-C. Nebel, and F. Flórez-Revuelta, “Recognition of activities of daily living with egocentric vision: A review,” *Sensors*, vol. 16, p. 72, 01 2016.

[27] D. Liang and E. Thomaz, “Audio-based activities of daily living (adl) recognition with large-scale acoustic embeddings from online videos,” *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 3, Mar. 2019.

[28] D. De, P. Bharti, S. Das, and S. Chellappan, “Multimodal wearable sensing for fine-grained activity recognition in healthcare,” *IEEE Internet Computing*, vol. 19, pp. 1–1, 09 2015.

[29] P. Urwyler, L. Rampa, R. Stucki, M. Büchler, R. M. Müri, U. P. Mosimann, and T. Nef, “Recognition of activities of daily living in healthy subjects using two ad-hoc classifiers,” in *Biomedical engineering online*, 2015.

[30] R. Stucki, P. Urwyler, L. Rampa, R. Müri, U. Mosimann, and T. Nef, “A web-based non-intrusive ambient system to measure and classify activities of daily living,” *Journal of medical Internet research*, vol. 16, p. e175, 07 2014.

[31] J. W. Ratcliff and D. Metzener, “Pattern matching: The gestalt approach,” *Dr. Dobb’s Journal*, July 1988.

[32] N. C. Krishnan and D. J. Cook, “Activity recognition on streaming sensor data,” *Pervasive and mobile computing*, vol. 10, pp. 138–154, 2014.

[33] S. Enshaeifar, A. Zoha, A. Markides, S. Skillman, S. T. Acton, T. Elsaleh, M. Hassanpour, A. Ahrabian, M. Kenny, S. Klein, H. Rostill, R. Nilforooshan, and P. Barnaghi, “Health management and pattern analysis of daily living activities of people with dementia using in-home sensors and machine learning techniques,” *PLOS ONE*, vol. 13, pp. 1–20, 05 2018.

[34] M. Van den Bergh, X. Boix, G. Roig, B. de Capitani, and L. Van Gool, "Seeds: Superpixels extracted via energy-driven sampling," in *Computer Vision – ECCV 2012* (A. Fitzgibbon, S. Lazebnik, P. Perona, Y. Sato, and C. Schmid, eds.), pp. 13–26, 2012.

[35] A. Banerjee, D. Maas, M. Bocca, N. Patwari, and S. Kasera, "Violating privacy through walls by passive monitoring of radio windows," *ACM Conference on Security and Privacy in Wireless & Mobile Networks*, pp. 69–80, 2014.

[36] P. Mohan, B. Lee, T. Chaspari, and C. R. Ahn, "Capturing occupant routine behaviors in smart home environment using hierarchical clustering models," *Construction Research Congress 2020*, March 2020.

[37] M. Gygli, H. Grabner, H. Riemenschneider, and L. Van Gool, "Creating summaries from user videos," in *Computer Vision – ECCV 2014*, pp. 505–520, Springer International Publishing, 2014.

[38] X. Han, Z. Li, H. Huang, E. Kalogerakis, and Y. Yu, "High-resolution shape completion using deep neural networks for global structure and local geometry inference," in *Proceedings of the IEEE International Conference on Computer Vision*, pp. 85–93, 2017.

[39] P. Felzenszwalb and D. Huttenlocher, "Efficient graph-based image segmentation," *International Journal of Computer Vision*, vol. 59, pp. 167–181, 09 2004.